LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

numero 201
maggio 1985
anno XVIII
volume XXXIV

L'ottemperanza agli accordi sul controllo delle armi

Le attività militari in Unione Sovietica possono essere controllate unilateralmente dagli Stati Uniti con un'ampia gamma di tecnologie di telerilevamento tra cui le fotografie ad alta risoluzione riprese da satellite

di David Hafemeister, Joseph J. Romm e Kosta Tsipis

La recente decisione degli Stati Uniti e dell'Unione Sovietica di riprendere i negoziati bilaterali sul controllo degli armamenti è stata annunciata da entrambi i governi con dichiarazioni pubbliche nelle quali si dava particolare rilievo alla necessità di fermare e addirittura di invertire la crescita dei rispettivi arsenali nucleari e dei relativi sistemi di armi. Dato il clima di sfiducia che regna oggi fra le due nazioni, è imperativo che l'ottemperanza ai termini di qualsiasi trattato che possa uscire dai nuovi negoziati sia controllabile unilateralmente da ognuna delle due parti. Il controllo esige che ogni nazione disponga di mezzi sicuri e obiettivi per seguire passo passo le attività militari dell'altra. Nel corso degli anni, pertanto, entrambi i paesi hanno messo a punto una famiglia completamente nuova di sistemi intesi a raccogliere tali informazioni a distanza. Questi sistemi vengono indicati collettivamente come mezzi tecnici nazionali di controllo.

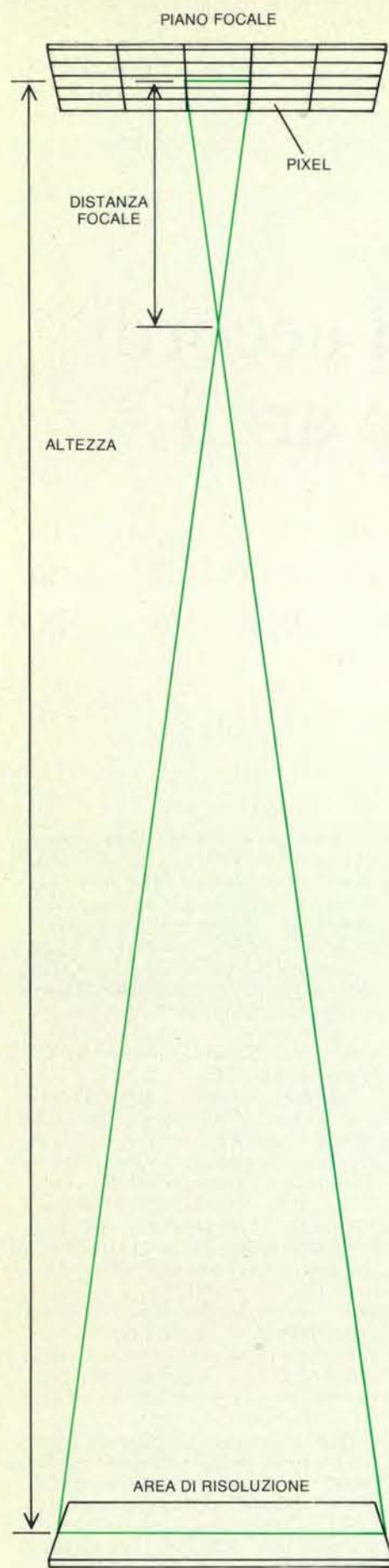
La questione del controllo è completamente disgiunta non solo dal problema legale che si pone quando si tratta di stabilire se una determinata attività costituisce una violazione del trattato, ma anche dal problema politico di che cosa fare a proposito di una violazione una volta che essa sia stata rilevata. Ciononostante, il controllo è qualcosa di più di un semplice problema tecnico. Negli Stati Uniti, per esempio, esso è stato uno dei punti centrali del ricorrente dibattito politico sui meriti dei vari trattati, esistenti e proposti, per il controllo degli armamenti. Gli oppositori di un qualsiasi trattato tendono a sostenere che, attenendosi ai termini di un trattato, gli Stati Uniti non possono controllare l'ottemperanza sovietica tanto bene da riuscire a garantire la sicurezza nazionale, mentre i fautori dei trattati tendono a sostenere che gli Stati Uniti sono in grado di farlo.

In genere la questione della controllabilità o meno di un trattato da parte degli Stati Uniti si può ridurre a due problemi più ristretti. Primo, a quale livello di attività clandestina sovietica sarebbe messa in pericolo la sicurezza degli Stati Uniti? Secondo, il sistema operativo di controllo degli Stati Uniti è in grado di rilevare quel livello di attività?

Quali sono le tecnologie necessarie per osservare lo sviluppo, la sperimentazione, la produzione e lo spiegamento di un sistema di armi? Per quel che riguarda lo sviluppo, i primi segni individuabili potrebbero essere costituiti dalle comunicazioni tra i funzionari e gli scienziati che lavorano sull'arma in questione; alcuni di questi segnali potrebbero essere intercettati per mezzo di onde radio di controllo. Oppure, l'arma o i suoi piani potrebbero essere osservati direttamente da una spia. Come ebbe a dire nel 1979 William J. Perry, a quel tempo sottosegretario alla Defense per la ricerca e la tecnologia, «noi seguiamo tanto bene l'attività a livello di ufficio progetti [dei missili sovietici] da essere sempre riusciti a prevedere ogni ICBM [missile balistico intercontinentale] prima ancora che venissero iniziate le prove ... Già in passato siamo stati in grado di scoprire l'esistenza di un ICBM prima della fase sperimentale».

Una volta che un'arma arriva alla fase di prova fuori del laboratorio è possibile servirsi di tutto lo spettro di energia acustica ed elettromagnetica per ottenere informazioni al riguardo. Rivelatori nella luce visibile, come quelli installati sui satelliti per la ricognizione fotografica, possono «vedere» l'arma direttamente; rivelatori nell'infrarosso installati sui satelliti possono «percepire» il calore emesso dal pennacchio di scarico di un razzo durante gli esperimenti; il radar può inseguire un'arma in volo, il sonar può inseguirla in acqua e i sismografi possono rilevare e valutare un esperimento nucleare sotterraneo.

Dopo essere stata sperimentata in tutto e per tutto, un'arma passa alla produzione. Questa comporta il trasporto in grande scala da e per gli impianti industriali, un'attività che può essere controllata servendosi delle radiazioni comprese nella regione dell'infrarosso o del vi-

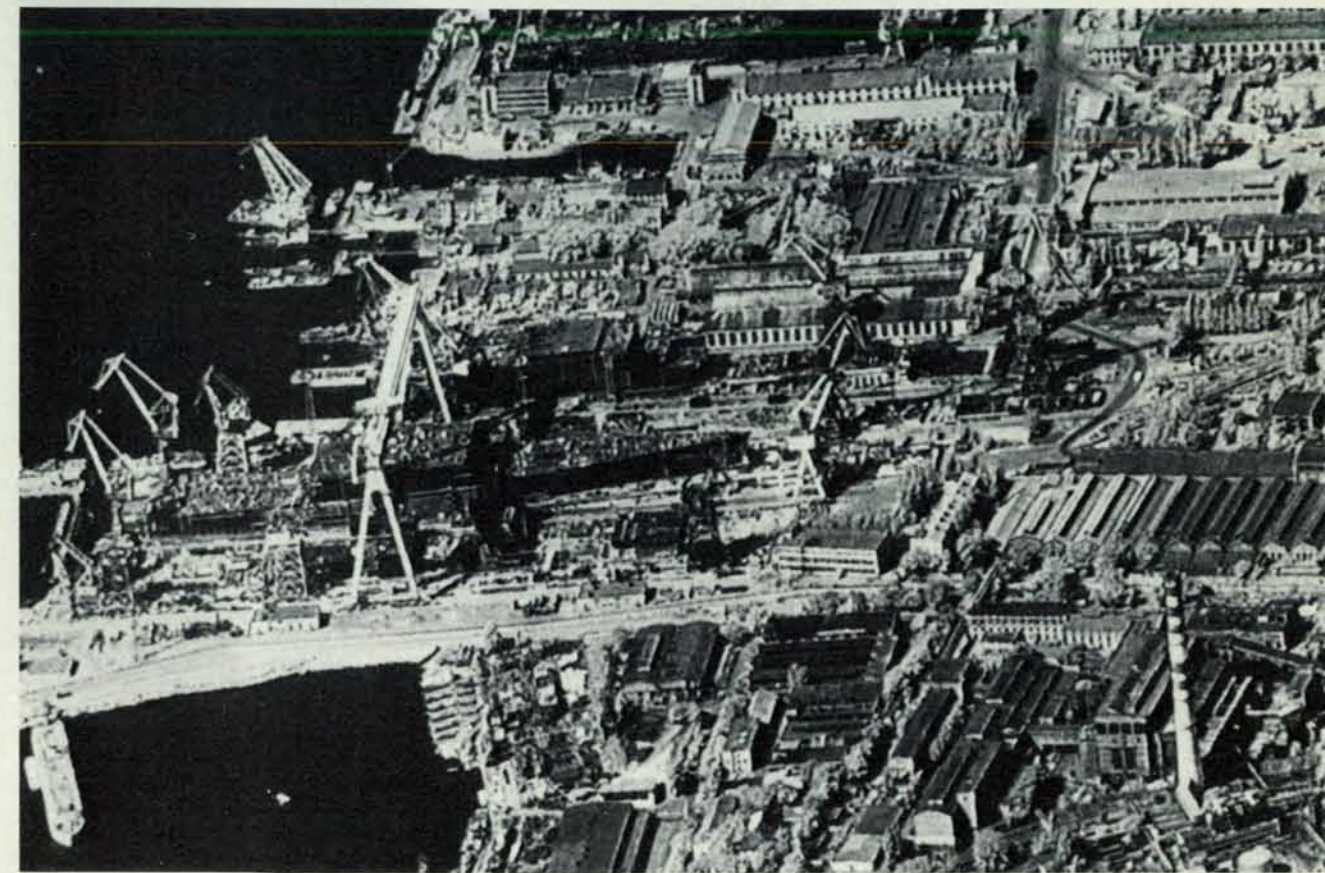
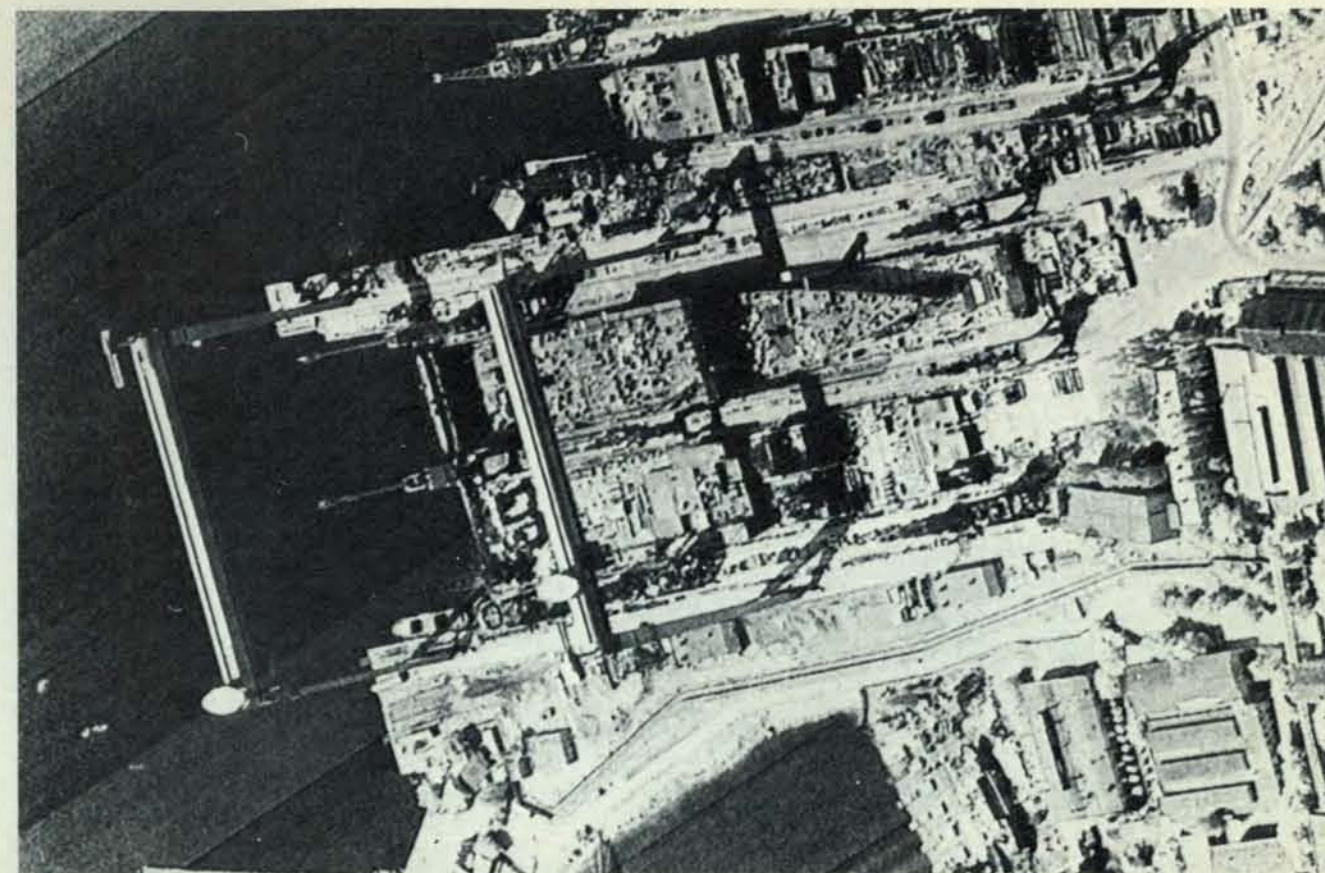
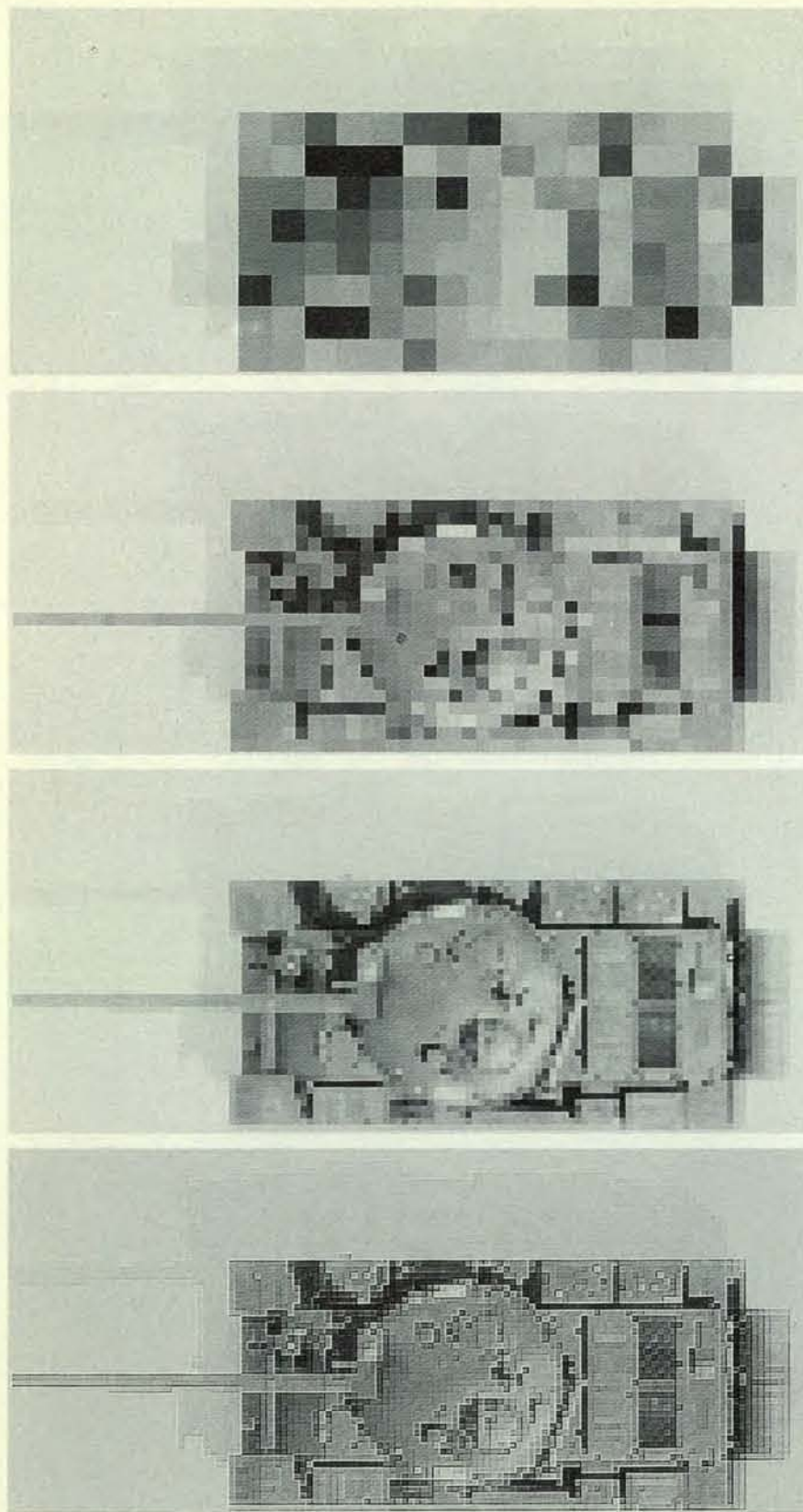
sibile. Infine è possibile seguire anche lo spiegamento operativo e l'addestramento degli addetti alla nuova arma. La principale differenza esistente nell'osservazione di questi processi è costituita dal fatto che, in linea di massima, la sperimentazione e la produzione debbono per forza essere rilevate mentre sono in corso, mentre lo spiegamento comporta attività che vanno avanti per anni e che possono essere rilevate in qualsiasi momento dopo la fase di avvio.

Per stabilire se l'Unione Sovietica sta violando i termini di un trattato non è però necessario osservare tutto ciò che ha a che fare con il laborioso processo che culmina con lo spiegamento dell'arma. Al contrario, basterebbe conoscere in modo sufficientemente particolareggiato solo alcune parti dell'attività in questione per individuarne l'intento e rilevare la violazione dell'accordo. Questo processo di rilevamento e identificazione è facilitato in molti modi dal sinergismo tra i vari mezzi impiegati per raccogliere informazioni dall'interno dell'Unione Sovietica. Le informazioni raccolte per mezzo di un certo sistema operativo di controllo (per esempio, l'intercettazione elettronica di messaggi) può dire a un altro sistema (come un satellite per la ricognizione fotografica) dove guardare e che cosa cercare. Frammenti di informazioni di per sé non decisivi, captati da un sistema, possono inoltre rivelare, una volta che sono interpolati e correlati con altre informazioni frammentarie raccolte da altri sistemi operativi di controllo, la natura di un'attività in via di svolgimento.

Mettere assieme i vari pezzi del *puzzle* è importante quanto raccoglierli. A questo compito è intrinseca comunque una certa dose di ambiguità e di incertezza. Alcune attività rilevate potrebbero non essere osservate in modo abbastanza particolareggiato da permettere di stabilire se costituiscono o meno una violazione delle clausole di questo o quel trattato. Altre attività potrebbero essere rilevate con dovizia di particolari, ma le relative clausole del trattato potrebbero essere troppo ambigue per permettere di stabilire se le attività in questione costituiscono o non costituiscono delle violazioni. Gli sforzi della comunità dei tecnici sono stati dedicati a ridurre queste incertezze mediante lo sviluppo di tecnologie di sorveglianza che dei segnali intercettati producono immagini e registrazioni estremamente particolareggiate prive al massimo di qualsiasi disturbo o rumore.

Il potere risolutivo al suolo di un satellite per la ricognizione fotografica dipende dalla dimensione dell'«area di risoluzione». Il diametro di una singola area di risoluzione della scena al suolo (*r*) è dato dalla formula $r = (h/f)d$, dove *h* è l'altezza del satellite, *f* la distanza focale del sistema ottico e *d* il diametro di un pixel, o elemento d'immagine, nel mezzo di registrazione. (Tutte le unità sono calcolate in genere in centimetri.) La dimensione di un pixel è determinata o dalla granulosità della pellicola fotografica o dalle dimensioni di una singola cella del dispositivo a scorrimento di carica (CCD) usato sul piano focale del sistema ottico per registrare l'immagine. L'area di risoluzione della generazione attuale di satelliti per la ricognizione fotografica sarebbe all'incirca di 10^2 centimetri.

Fatti e attività in via di svolgimento che potrebbero provocare nell'ambiente fisico cambiamenti permanenti visibili vengono individuati in modo più attendibile per mezzo della ricognizione fotografica effettuata da satelliti, cioè per mezzo del periodico fotografare di scene all'interno dell'Unione Sovietica mediante sistemi ottici installati a bordo di satelliti immessi in un'orbita relativamente bassa. Fotografando a più riprese una scena nelle stesse condizioni, si è in grado di stabilire se in essa sono intervenuti dei cambiamenti.

Dato un numero sufficiente di particolari, è possibile riconoscere l'attività che ha causato il cambiamento. Un tipo speciale di macchina fotografica raccoglie la luce riflessa dalla scena e forma un'immagine costituita da un insieme di punti chiari e scuri (o in colore) su una superficie fotosensibile di registrazione. Ogni punto è un *pixel*, ossia un elemento di immagine. Le dimensioni dei pixel, la distanza focale della macchina fotografica e la quota a cui si trova il satellite determinano il potere risolutivo del sistema, che è la dimensione dell'oggetto più piccolo al suolo che il sistema stesso è in grado di distinguere (*si veda l'illustrazione in questa pagina*). Quanto più fini sono i particolari visibili in una immagine, tanto più facile è rilevare i cambiamenti intervenuti da una fotografia all'altra e tanto più piccoli sono i cambiamenti che si possono individuare. L'aumento di risoluzione accresce il contenuto di informazioni delle immagini. Il maggior contenuto di informazioni fa aumentare le probabilità di individuare anche l'attività responsabile dei cambiamenti osservati.

L'esperienza ha dimostrato che si può rilevare la presenza di un oggetto in una scena se l'oggetto è grande almeno quanto l'area di risoluzione. Se è otto volte più grande, può essere riconosciuto (poniamo come un carro armato o come un autocarro) e se è dodici volte più grande può essere addirittura identificato (come un vecchio carro armato sovietico T-62 o come un più recente T-72). Si dice che l'area di risoluzione dei satelliti da ricognizione degli Stati Uniti sia inferiore a un quadrato di 10 centimetri di lato. Si può supporre quindi che tali sistemi siano in grado di rilevare al suolo un oggetto più o meno delle stesse dimensioni e di identificarne perfettamente uno che abbia un diametro inferiore a un metro e mezzo. L'identificazione di oggetti e di attività diventa ancora più facile se si dispone di immagini fotogra-

Le fotografie da satellite intensificate al calcolatore rivelano la prima portaerei a propulsione nucleare della marina sovietica in costruzione in un cantiere del Mar Nero. La nave da guerra da 75 000 tonnellate, che si suppone verrà chiamata *Cremlino* quando diventerà operativa nel 1994, viene costruita in due sezioni: la sezione di prora, lunga 264 metri, e la sezione di poppa, lunga 73 metri, visibili in due scali adiacenti sotto la gigantesca gru a cavalletto. Le fotografie sono state pubblicate per la prima volta nel 1984 su «Jane's Defence Weekly».

La sequenza di immagini del plastico di un carro armato sovietico simula l'effetto dell'aumento di risoluzione. Una fotografia del modello è stata sottoposta a scansione ottica in modo da produrre una registrazione digitale su nastro magnetico. Le informazioni sul nastro sono state poi elaborate così da riprodurre l'immagine a tre diverse dimensioni di pixel, corrispondenti grosso modo a tre risoluzioni. Alla prima risoluzione l'oggetto può essere identificato come un autocarro (*in alto*); alla seconda come un carro armato (*seconda immagine dall'alto*); alla terza come un T-62 sovietico (*seconda immagine dal basso*). L'immagine intensificata in basso è stata sottoposta a speciale elaborazione per definirne meglio i bordi e regolarne il contrasto.

fiche formate da radiazione riflessa di lunghezze d'onda diverse. Combinando le immagini di una scena ripresa nell'infrarosso e nell'ultravioletto oltre che nel visibile, un analista può dedurre ulteriori informazioni su un oggetto o individuarne altri camuffati.

Una volta formata dalla macchina fotografica, l'immagine al suolo deve essere registrata e trasmessa per l'analisi e l'interpretazione alle attrezzature a terra. La registrazione può essere fatta o su pellicola fotografica o su una schiera bidimensionale di rivelatori elettroottici fotosensibili, chiamati dispositivi a scorrimento di carica (o CCD da *charge-coupled device*). Un rivelatore di questo tipo trasforma la quantità di luce ricevuta in un breve e determinato periodo di tempo in una quantità proporzionale di carica elettrica. In questo modo lo schema luminoso crea sulla schiera dei rivelatori una copia elettrica di se stesso, che a sua volta viene convertita in una sequenza di numeri, la quale è poi trasmessa a un ricevitore a terra. Le attrezzature della stazione a terra trasformano in una fotografia la copia elettrica dello schema luminoso che è stato registrato dalla schiera dei rivelatori. Il processo viene poi ripetuto e viene registrata una nuova immagine.

Nei sistemi con una grande distanza focale le schiere elettroottiche possono raggiungere dimensioni di pixel paragonabili alla migliore pellicola fotografica. Esse presentano inoltre numerosi vantaggi rispetto alle pellicole. Le immagini registrate su pellicola possono essere trasmesse nell'uno o nell'altro di due modi: si può sviluppare la pellicola a bordo del satellite in modo analogo a quello in cui opera una Polaroid, e in tal caso l'immagine viene trasmessa successivamente con un sistema simile a una telecamera; oppure la pellicola può essere espulsa dal satellite dentro capsule speciali, le quali rientrano nell'atmosfera e, grazie a un paracadute, scendono lentamente fino a essere prese a mezz'aria da un aereo particolarmente attrezzato. Entrambi questi metodi provocano ritardi nel ricevimento delle immagini e limitano la vita utile del satellite, in quanto alla fine gli apparecchi fotografici rimangono a corto di pellicola. Con l'ausilio di un satellite ripetitore, i rivelatori fotosensibili installati a bordo dei satelliti statunitensi trasmettono in tempo reale, ossia mentre le osservano, le scene che sorvolano. Le immagini ricreate a terra hanno una gamma dinamica superiore di molto a quella ottenuta per mezzo di pellicola fotografica; di conseguenza non sono altrettanto sensibili a grandi cambiamenti nell'intensità dell'illuminazione.

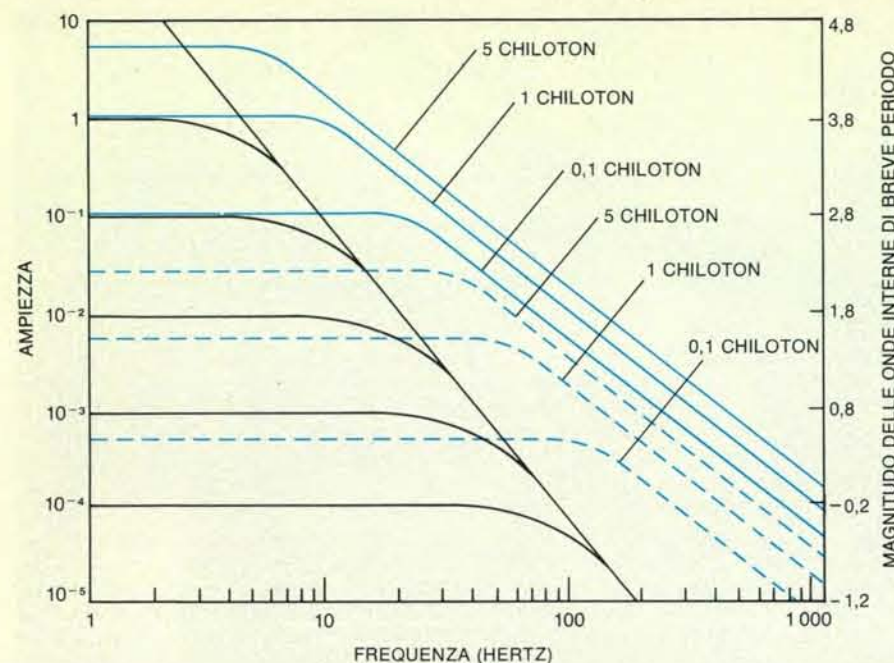
Dopo che l'immagine di una scena è stata ricevuta da una stazione a terra è possibile accrescerne la qualità ottica per mezzo di calcolatori veloci e di dispositivi di elaborazione specializzati.

L'immagine radar di un grande vortice nel Golfo del Messico rivela la scia di una nave (*in basso*). Variazioni nella irregolarità della superficie associata al vortice e alla scia sono state registrate dal radar ad apertura sintetizzata a bordo del satellite *Seasat*. L'elaborazione dell'immagine è stata effettuata al Jet Propulsion Laboratory del California Institute of Technology. Il vantaggio principale dei sistemi radar a bordo di satelliti consiste nel fatto che essi funzionano in qualsiasi ora del giorno o della notte e indipendentemente dalle condizioni meteorologiche.

La tecnica è nota come elaborazione digitale delle immagini. È possibile per esempio eliminare da una fotografia la sfocatura delle linee e delle forme causata dalla turbolenza, dalla mancata compensazione del moto, dalla sovraesposizione e dallo scarso contrasto alterando artificialmente le zone di grigio e accentuando quelle chiare e quelle scure. L'elaborazione delle immagini permette di identificare forme caratteristiche (come quella di un silo missilistico visto dall'alto) in fotografie del terreno o di completare una scena un poco coperta dalle nuvole.

La capacità dei satelliti per la ricognizione fotografica di raccogliere informazioni è limitata sia dal fatto che un sistema a luce visibile non può funzionare di notte, sia dal fatto che le radiazioni visibile, ultravioletta e infrarossa non possono penetrare una coltre di nubi. Pertanto il cattivo tempo e la notte polare potrebbero impedire ai satelliti statunitensi per la ricognizione fotografica di osservare estese zone dell'Unione Sovietica per periodi di tempo prolungati. Questo problema è stato risolto ricorrendo a un'altra parte dello spettro elettromagnetico: le onde radio.

Avendo una lunghezza d'onda superiore a quella della radiazione visibile, le onde emesse da un sistema radar non vengono modificate in misura significativa da una coltre di nubi o dalla pioggia. Pertanto possono «illuminare» il loro bersaglio indipendentemente dalle condizioni meteorologiche. Poiché generano la propria radiazione riflessa, i radar possono funzionare a qualsiasi ora del giorno. Il potere risolutivo di un sistema radar è proporzionale alla lunghezza d'onda della radiazione che esso emette e inversamente proporzionale alla lunghezza della sua antenna: per questa ragione la risoluzione delle immagini radar del terreno è normalmente molto bassa. Si può aumentare il grado di risoluzione facendo sì che alle onde radar riflesse l'antenna piuttosto piccola del satellite appaia molto lunga. Ciò si ottiene con un metodo noto come radar ad apertura sintetizzata (SAR). Un satellite SAR combina la minuziosa elaborazione elettronica delle onde riflesse dall'obiettivo con il moto relativo del satellite rispetto alla superficie della Terra per ottenere immagini radar ad alta risoluzione (si veda l'articolo *Immagini radar della Terra dallo spazio* di Charles Elachi in «Le Scienze» n. 174, febbraio 1983). La risoluzione di un'immagine SAR non è elevata quanto quella di un'immagine analoga ripresa in luce visibile: c'è un limite pratico alle dimensioni alle quali un'antenna «virtuale» può apparire alle onde radio. Essa per altro è abbastanza buona da permettere agli Stati Uniti di seguire molte attività in aree dell'Unione Sovietica coperte da nubi o illuminate soltanto per qualche ora al giorno.

Non è necessario che i radar siano a bordo di satelliti per fornire informazioni utili. I radar con base a terra, su nave o su aereo sono i mezzi primari con i quali gli Stati Uniti ottengono informazioni sui missili sovietici mentre vengono sperimentati. Questa copertura è un buon esempio del sinergismo creato dalla molteplicità di attrezzature operative di controllo degli Stati Uniti. Il lancio sperimentale di missili balistici sovietici, che ha luogo nelle basi di lancio di Plesetsk e Tyuratam in Asia centrale, è preceduto da molte attività. Per esempio, il primo allarme di un esperimento imminente potrebbe venire da un satellite per la ricognizione fotografica che ha ripreso il missile mentre viene preparato sulla rampa di lancio. Il segnale per attivare tutti i sensori potrebbe venire da comunicazioni intercettate o da satelliti da avvistamento avanzato, immessi in orbite geostazionarie sopra l'equatore: questi satelliti «fissano» in continuazione l'Unione Sovietica e possono rilevare la radiazione infrarossa emessa dal pennacchio infuocato dei gas di scarico di un razzo non appena il missile esce dalla

La capacità di distinguere le esplosioni nucleari sotterranee dai terremoti si estende ormai fino a potenze esplosive dell'ordine di un chiloton o anche meno, come risulta da questo grafico, basato sul lavoro di Jack F. Evernden dell'US Geological Survey. Le curve in nero rappresentano i terremoti; quelle continue in colore esplosioni nucleari in roccia tenera e le tratteggiate esplosioni nucleari di cui viene ridotto l'effetto disponendo l'esplosivo in cavità di miniera.

coltre di nubi sopra il sito di lancio. A quel punto il mezzo principale per raccogliere informazioni diventa il rilevamento delle onde radio.

Grandi radar con base a terra seguono il moto del missile durante i vari stadi della fase di spinta. Il radar può misurare continuamente e con grande precisione la velocità del missile e quindi la sua accelerazione misurando lo spostamento Doppler dell'onda riflessa. Quando il sito di lancio di un missile è troppo all'interno del territorio sovietico per essere osservato dai radar convenzionali con base a terra, gli Stati Uniti si servono di radar con portata «oltre l'orizzonte». Questi dispositivi proiettano fasci che la ionosfera, facendo da specchio, riflette all'interno dell'Unione Sovietica. Un radar di questo tipo fornisce informazioni utili anche sulla velocità di un missile.

Radioricevitori sia a terra sia su satelliti in orbite alte intercettano il flusso di messaggi che i missili rimandano a terra durante gli esperimenti e che descrivono le prestazioni e le condizioni delle varie parti del missile. Questi messaggi, noti collettivamente come telemetria, comprendono informazioni sulla quantità di combustibile bruciata, sugli ordini che il sistema di guida sta impartendo al razzo e sulla temperatura e sulla pressione alle quali sono sottoposte le varie parti del missile. Tutti questi segnali, necessari ai tecnici per controllare se il missile funziona secondo le previsioni, sono intercettati anche dalle attrezzature di controllo statunitensi. Correlando questi dati telemetrici con informazioni sul moto del missile raccolte dai radar a terra, i funzionari del servizio informazioni statunitense sanno non solo che un missile sovietico è stato sperimentato ma, in gran parte, anche quali sono le sue caratteristiche operative. Alcuni segnali telemetrici relativi a questo tipo di esperimenti sono stati cifrati dall'URSS allo scopo di renderli incomprensibili agli analisti americani, una pratica contro la quale gli Stati Uniti hanno protestato sostenendo che in base al trattato, non ratificato, SALT II gli unici segnali che possono essere cifrati sono quelli non direttamente attinenti al controllo del trattato. Anche senza accesso alla telemetria non cifrata, però, gli Stati Uniti possono rilevare con estrema attendibilità l'esperimento missilistico sovietico in questione e possono anche determinare certe importanti caratteristiche operative di un missile balistico, come il numero dei veicoli di rientro che sta trasportando.

La fase finale dell'esperimento di un missile balistico è costituita dal ritorno dei suoi veicoli di rientro nell'atmosfera e dal loro impatto al suolo. Dal momento che i fattori atmosferici influiscono sulla traiettoria di volo dei veicoli di rientro, informazioni sulle loro prestazioni in questa fase terminale forniscono una misura importante della precisione del missile e la loro acquisizione è cruciale per l'Unione Sovietica. Per questo motivo, durante gli esperimenti missilistici sovietici, i veicoli di rientro vengono diretti in attrezzate zone d'impatto o nella penisola di Kamčatka o nell'oceano Pacifico centrale. Gli Stati Uniti hanno sviluppato tutta una gamma di sistemi di rilevamento che seguono il ritorno dei veicoli di rientro sovietici in queste aree. Radar a schiera in fase molto precisi (in grado di individuare a migliaia di chilometri di distanza oggetti grandi quanto un pallone da pallacanestro e di inseguirne simultaneamente centinaia), radioricevitori, telescopi per infrarosso e per luce visibile equipaggiati con macchine fotografiche e spettrometri a scansione rapida registrano con dovizia di particolari le varie forme di energia radiante emessa e riflessa dal veicolo di rientro. Alcuni di questi strumenti sono installati nell'isola di Shemya alla punta estrema delle Aleutine e sull'atollo di Kwajalein nel Pacifico; altri sono trasportati da navi e da aerei e perfino da piccoli razzi sonda lanciati prima dell'arrivo del veicolo di rientro.

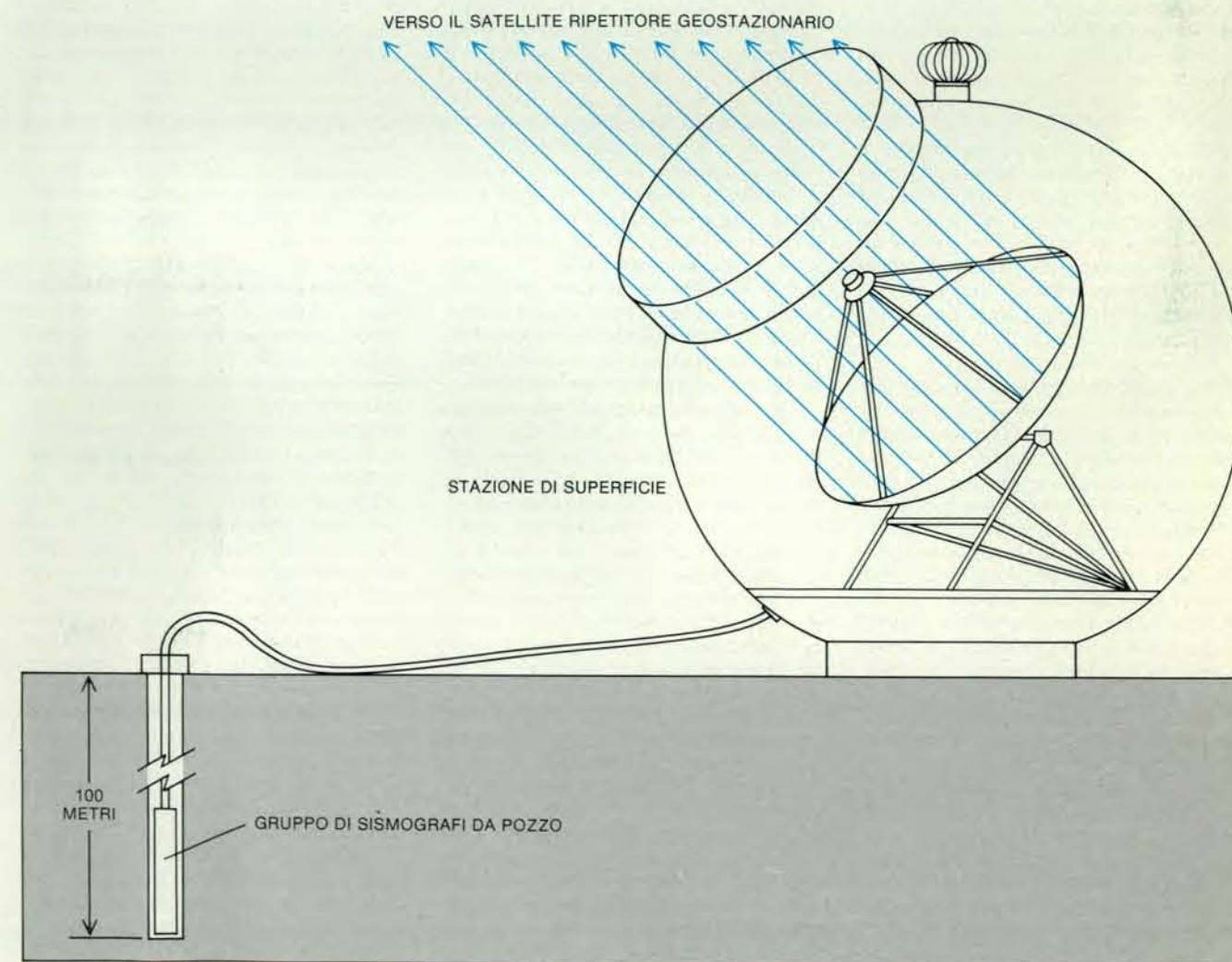
Da tutti questi dati, correlati e interpretati dagli analisti del servizio informazioni, gli Stati Uniti possono valutare con un alto grado di attendibilità molte delle caratteristiche dell'arma sovietica sperimentata. (Per esempio, quanto è grande il razzo vettore? Qual è il raggio d'azione? Quanto pesano i veicoli di rientro?) Altre caratteristiche sono molto più difficili da determinare. La precisione, per esempio, può essere stabilita solo statisticamente dalla traiettoria di volo e dal punto d'impatto dei veicoli di rientro. Se gli Stati Uniti non conoscono esattamente in quale punto i sovietici hanno indirizzato i loro veicoli di rientro, nelle loro stime sulla precisione dei veicoli si introduce un notevole grado di incertezza. Anche la valutazione dell'affidabilità degli ICBM e dei veicoli di rientro sovietici è incerta dal punto di vista statistico. Alcune caratteristiche sono estremamente difficili da determinare. La potenza esplosiva potenziale dell'arma che verrà messa sul veicolo di rientro può essere dedotta soltanto con grande approssimazione dal peso del veicolo. Ciononostante, si può fare sicuro affidamento sui mezzi tecnici unilaterali americani di controllo degli esperimenti con missili balistici sovietici non soltanto per rilevare questo tipo di esperimenti, ma anche per fornire agli Stati Uniti moltissime informazioni sulle prestazioni dell'arma sperimentata.

Le testate nucleari che i missili trasportano debbono essere messe a punto per mezzo di esperimenti sotterranei. Questo stato di cose è la conseguenza del Limited Test Ban Treaty, firmato dagli Stati Uniti, dalla Gran Bretagna e dall'Unione Sovietica nel 1963 e che bandiva tali esperimenti dall'atmosfera. La tecnologia di rilevamento necessaria per controllare gli esperimenti sotterranei si basa su sismografi che rilevano le onde acustiche generate da un'esplosione, su attrezzature di registrazione e su calcolatori che analizzano i dati. Se venissero ripresi, gli esperimenti nell'atmosfera sarebbero controllati facilmente da satelliti speciali e da tutto un insieme di altre tecniche, fra cui molte di quelle di cui abbiamo parlato.

Le esplosioni sotterranee generano onde elastiche che si propagano per lunghissime distanze sia in superficie sia attraverso la crosta terrestre. La magnitudo e il punto di origine delle onde elastiche possono essere determinati con l'ausilio di schiere di sismografi sensibili. I sismografi sono «accesi» permanentemente e quindi rilevano e registrano tutte le scosse telluriche, comprese quelle causate da un'esplosione.

Un problema persistente nel rilevamento sismico delle esplosioni nucleari sperimentali sotterranee era costituito dal fatto che tali esplosioni producono onde elastiche quasi analoghe a quelle di molti altri eventi naturali, come i terremoti. Esisteva quindi il rischio di scambiare un'esplosione nucleare per un terremoto e viceversa. Ricerche successive hanno eliminato questa difficoltà salvo per le esplosioni nucleari più piccole (si veda l'articolo *Il controllo di un bando totale agli esperimenti nucleari* di Lynn R. Sykes e Jack F. Evernden in «Le Scienze» n. 172, dicembre 1982). Rimanevano però due possibili sorgenti di evasione. L'Unione Sovietica avrebbe potuto predisporre di sperimentare una piccola esplosione nucleare durante un terremoto: le onde del terremoto avrebbero così potuto coprire il segnale generato dalla detonazione. Per un certo tempo sembrò anche che l'energia di un ordigno di bassa potenza esplosiva potesse essere camuffata efficacemente facendo esplodere l'arma in una grande cavità sotterranea: in questo modo l'esplosione sarebbe sfuggita a qualsiasi rilevamento al di fuori dell'Unione Sovietica. Di recente però si è avuta notizia che un gruppo di geofisici dell'ufficio dell'US Geological Survey di Menlo Park, in California, ha messo a punto un modo per rilevare in maniera inequivocabile, con una migliore attrezzatura operativa di controllo, una piccola esplosione nucleare dell'ordine di un chiloton o anche meno.

La nuova tecnica di rilevamento si avvale del fatto che un terremoto è un evento molto esteso: di conseguenza l'energia acustica che esso emette è concentrata a grandi lunghezze d'onda. Un'esplosione nucleare, d'altra parte, è un evento puntiforme che libera repentinamente nell'ambiente una quantità enorme di energia. Tanto l'estensione limitata quanto il carattere repentino di tale evento fanno sì che esso liberi la propria energia in onde molto più piccole e quindi di frequenza più elevata. Di conseguenza le onde dovute a un'esplosione nucleare possono essere rilevate e riconosciute per le loro caratteristiche di frequenza anche se sono accompagnate da quelle di un terremoto.

Le onde di un terremoto, per esempio, possono coprire il segnale dovuto all'esplosione di un chiloton, di cui si è cercato di ridurre l'effetto, fino a frequenze di circa 10 hertz (*si veda l'illustrazione nella pagina precedente*). Le onde sismiche, tuttavia, non contengono energia

La stazione sismica senza personale messa a punto ai Sandia National Laboratories potrebbe controllare l'ottemperanza a un trattato che preveda il bando totale agli esperimenti nucleari. Un sistema di cinque stazioni di questo genere è in funzione negli Stati Uniti e in Canada.

che sia rilevabile nella banda di frequenze al di sopra di circa 30 hertz, mentre le onde dovute a un'esplosione contengono energia rilevabile fino a frequenze di parecchie centinaia di hertz. Per questo motivo i sismografi sintonizzati in modo da rilevare soltanto onde di alta frequenza non «sentiranno» i terremoti; essi non verranno confusi dal rumore costante della Terra, ma rileveranno facilmente le onde generate dall'esplosione.

Poiché le onde acustiche di alta frequenza non percorrono lunghe distanze attraverso la crosta terrestre, il loro rilevamento richiederebbe l'installazione di sismografi all'interno dell'Unione Sovietica. Questo problema politico potenzialmente delicato è stato risolto dalla messa a punto di una stazione sismica senza personale. Ai Sandia National Laboratories un gruppo di ricercatori ha ideato, sperimentato e messo in opera negli Stati Uniti e nel Canada, per far pratica e a fini dimostrativi, cinque stazioni sismiche senza personale (*si veda l'illustrazione nella pagina precedente*). Queste stazioni comunicano costantemente mediante collegamento via satellite con un impianto centrale di controllo negli Stati Uniti, il che le rende praticamente a prova di manomissione. Durante le trattative per un bando totale agli esperimenti, che furono interrotte dagli Stati Uniti nel 1980 dopo l'invasione sovietica dell'Afghanistan, l'Unione Sovietica ha accettato una proposta per la collocazione di stazioni di questo tipo sul suo territorio. Sembra pertanto che non vi siano ostacoli di ordine tecnico alla possibilità di controllare con grande sicurezza esperimenti nucleari sotterranei che liberano quantità estremamente ridotte di energia dell'ordine di un chiloton o meno.

I tre tipi di mezzi tecnici nazionali che permettono di controllare le attività sovietiche di cui abbiamo parlato - quelli che creano immagini, quelli che rilevano le onde elettromagnetiche emesse o riflesse dai missili durante gli esperimenti e quelli che «percepiscono» le onde acustiche - esemplificano le raffinate capacità operative di controllo degli Stati Uniti. Anzi, la dovizia di particolari che gli Stati Uniti sono in grado di rilevare suggerisce il tipo di clausole di un trattato che gli Stati Uniti possono e non possono controllare. Il punto è importante: il principio guida per l'accettazione di qualsiasi clausola di un trattato deve essere che gli Stati Uniti possano essere sicuri di scoprire eventuali violazioni che potrebbero minacciare la loro sicurezza nazionale. La possibilità che gli Stati Uniti rilevino tali violazioni dipende sia da ciò che essi considerano come una minaccia per la loro sicurezza nazionale, sia dalle proprietà fisiche del sistema d'arma o dall'attività militare limitata dal trattato.

Per fare un esempio, considerate le stazioni sismiche senza personale all'interno dell'Unione Sovietica, con quali clausole di un trattato che limiti gli esperimenti nucleari sotterranei i rilevatori sismici americani possono controllare l'ottemperanza? Abbiamo già dimostrato che con tali stazioni gli Stati Uniti possono rilevare esperimenti nucleari sotterranei di una potenza esplosiva dell'ordine di un solo chiloton. Questa capacità permette davvero agli Stati Uniti di aderire a un trattato che metta al bando tutti gli esperimenti nucleari sotterranei, fiduciosi che venga individuata qualsiasi violazione sovietica che metta in pericolo la sicurezza americana? L'Unione Sovietica potrebbe però tentare di compiere esperimenti molto al di sotto del livello americano di rilevamento, conducendo un esperimento clandestino, poniamo, di 100 tonnellate. I sovietici hanno già condotto, peraltro, centinaia di esperimenti di entità molto maggiore; per di più, un esperimento di 100 tonnellate sarebbe 1000 volte inferiore alla loro arma nucleare strategica più piccola e 250 000 volte inferiore a quella più grande. Molto probabilmente, dunque, essi non trarrebbero alcuna informazione utile da un esperimento di livello tanto basso, anche se si trattasse di un nuovo tipo di arma. È difficile quindi che un esperimento molto al di sotto del livello di rilevamento di un chiloton possa minacciare la sicurezza degli Stati Uniti. Per contro, qualsiasi esperimento che possa dare loro informazioni utili verrebbe certamente individuato. Pertanto il confronto tra ciò che gli Stati Uniti devono necessariamente rilevare in maniera attendibile e ciò che possono rilevare in maniera attendibile in esperimenti nucleari sotterranei porta alla conclusione che gli Stati Uniti potrebbero controllare abbastanza bene un bando totale agli esperimenti, così da salvaguardare la loro sicurezza nazionale.

Valutazioni analoghe si possono fare a proposito della capacità degli Stati Uniti di controllare gli esperimenti missilistici sovietici. Per esempio, questa capacità è all'altezza di controllare un accordo che limiti i miglioramenti da apportare alla precisione degli ICBM? Il determinare unilateralmente la precisione di un missile osservando gli esperimenti di un missile balistico è, nel migliore dei casi, statisticamente incerto, e quindi i miglioramenti in fatto di precisione ottenuti senza tenere conto di una limitazione sarebbero difficili da individuare. D'altra parte, se gli Stati Uniti decidono che è importante tenere sotto controllo la precisione sovietica, c'è un'altra possibile alternativa: mettere al bando ogni esperimento di missili balistici. Controllare che un esperimento di un missile balistico abbia avuto luogo è una cosa che si può fare con un alto grado di affidabilità. Di fatto, data la molteplicità di mezzi di cui gli americani dispongono per rilevare questi esperimenti, le possibilità che gli Stati Uniti rilevino una singola prova di volo sovietica sono certo superiori al 90 per cento. Per determinare con sicurezza la precisione di un nuovo missile sono necessari almeno venti voli di prova. Pertanto, anche se gli Stati Uniti possono essere sicuri di rilevare l'esperimento di un missile balistico sovietico soltanto nel 90 per cento dei casi, la probabilità che non rilevino uno dei 20 voli di prova è pari soltanto a uno su 100 miliardi di miliardi (10^{20}). In poche parole, un trattato che metta al bando completamente la sperimentazione di missili balistici potrebbe essere controllato con sicurezza. Inoltre, siccome gli Stati Uniti possono stabilire il numero dei veicoli di rientro che un missile balistico è in grado di portare sul bersaglio, un altro trattato controllabile sarebbe quello che limita le prove di volo a quei missili progettati per portare una sola testata.

Quando si passa poi alle possibilità di controllo della produzione e dello spiegamento di armi strategiche sovietiche, non ci sono molti dubbi sul fatto che i satelliti americani per la ricognizione fotografica possano rilevare e contare in modo certo quei grandi sistemi di vettori di armi strategiche che sono gli ICBM, i missili balistici lanciati da sommergibili (SLBM), i sommergibili e i bombardieri. Il problema chiave è questo: con quale precisione è necessario che questi sistemi di arma siano controllati in modo da garantirsi contro l'erosione della sicurezza nazionale degli Stati Uniti provocata da violazioni di trattati che ne limitino o ne proibiscano la produzione o lo spiegamento?

Mantenere segrete eventuali attività compiute in violazione a un trattato che metta al bando un intero sistema di armi è particolarmente difficile. Le attività di vasta scala inerenti allo sviluppo, alla sperimentazione, alla produzione e allo spiegamento di un numero significativo di qualsiasi arma di grandi dimensioni si individuano facilmente. Anzi, qualsiasi tentativo di nascondere tali attività sarebbe ostacolato dal fatto che i sovietici non sanno esattamente ciò che gli Stati Uniti sono in grado di rilevare. Quindi non sanno che cosa cercare di nascondere e non sanno neppure con quale accuratezza è necessario, da parte loro, tentare di occultare i particolari di mutamenti visibili che verrebbero provocati dalla produzione proibita di un'arma.

La fiducia nella possibilità di controllare i limiti numerici posti a un sistema di armi dipenderebbe dal numero delle armi permesse. In linea di massima è difficile avere fiducia nel controllo di accordi che consentano un numero ridotto di armi. Il bando totale a un sistema missilistico sarebbe più facile da controllare di un accordo che consenta, poniamo, 100 di tali missili a ognuna delle due parti. Se si partisse dal presupposto che i sovietici non ne avessero affatto, non appena i sistemi operativi di controllo americani ne individuassero uno, risulterebbe subito evidente che i sovietici hanno violato l'accordo. Se però l'accordo ne permettesse 100, sarebbe difficilissimo sapere se essi ne hanno 100 o 120. L'ottemperanza a eventuali accordi sul controllo degli armamenti che proibiscano completamente un sistema di armi, una pratica o un'attività è molto più facile da individuare e da controllare dell'ottemperanza ad accordi che consentano un numero ridotto di armi.

L'Unione Sovietica ha per altro decine di sommergibili e di bombardieri, un migliaio di SLBM e circa 1400 ICBM. Si dice spesso che gli Stati Uniti siano in grado di controllare queste cifre con un margine di errore del 10 per cento o meno (una pretesa credibile alla luce della precedente disamina sulle capacità di controllo americane). Questa stima implica che gli Stati Uniti siano in grado di controllare l'ottemperanza a precisi limiti numerici con un margine di incertezza di qualche sommergibile o bombardiere e forse di 100 SLBM o ICBM. Dato il numero elevato di armi di cui entrambe le potenze dispongono attualmente, questa precisione in fatto di controllo sembra più che sufficiente a garantire la sicurezza degli Stati Uniti nell'ambito di un trattato che limiti la produzione di armi strategiche.

Un caso speciale è il problema presentato dai missili da crociera, piccoli «velivoli» radiocomandati lunghi alcuni metri. In questo caso il principale fattore di complicazione è che lo stesso tipo di missile da crociera può essere adatto sia per il trasporto di una testata nucleare sia per il trasporto di una testata convenzionale. Il compito quindi di distinguere i missili da crociera che trasportano testate nucleari da quelli che trasportano testate convenzionali è di una difficoltà estrema. È probabile che qualsiasi trattato controllabile riguardante i missili da crociera debba considerare il numero totale concesso, senza distinzione tra testate nucleari e convenzionali.

Abbiamo parlato di una molteplicità di potenti tecnologie di telerilevamento, ma soltanto di alcune loro applicazioni in fatto di controllo. (Non si è detto nulla, per esempio, sul modo in cui queste tecnologie potrebbero servire a controllare un trattato riguardante le armi antisatellite.) Inoltre, le effettive capacità degli Stati Uniti di raccogliere informazioni segrete sono molto maggiori di quanto si sia detto in questa sede. Non solo le tecnologie di controllo sono troppe per poterle discutere in un articolo come questo, ma, fatto ancora più significativo, la natura del processo di raccolta delle informazioni è tale che molti dei metodi e delle fonti per mezzo dei quali gli Stati Uniti vengono a conoscenza delle azioni sovietiche sono classificati. Anche la nostra trattazione necessariamente limitata fa pensare che la ricca capacità di telerilevamento degli Stati Uniti sia in grado di controllare adeguatamente una vasta gamma di trattati.

Macchine che comprendono la voce

Sfruttando la potenza dei moderni calcolatori numerici, è possibile interpretare il segnale vocale in termini di elementi fonetici significativi del linguaggio, mediante tecniche di ricerca su grafi

di Roberto Pieraccini

Il linguaggio parlato, frutto di un processo cognitivo che ha permesso in millenni di evoluzione del genere umano la specializzazione di strutture cerebrali, ha raggiunto con l'*Homo sapiens* un così elevato grado di complessità da richiedere l'uso di diversi livelli di conoscenza: acustica, lessicale, sintattica, fino a giungere, con i concetti semantici e pragmatici, ai livelli più astratti e più lontani dal mezzo fisico che li trasporta, il segnale acustico emesso dall'apparato fonatorio. Ciononostante, ci appare così semplice e così naturale da usare che viene coinvolto nella maggior parte delle nostre azioni e risulta insostituibile nei rapporti con i nostri simili. Nessun altro mezzo di comunicazione può essere paragonato a esso in termini di efficacia e potenzialità.

In questi ultimi anni è sorta l'esigenza di utilizzare il linguaggio parlato per comunicare non più solamente con gli esseri umani, ma anche con quei sistemi che, a un ritmo incalzante, stanno entrando nella vita di tutti i giorni: i calcolatori elettronici. Si è sviluppata così una nuova branca della scienza dei calcolatori, chiamata «riconoscimento automatico della voce» (ASR dall'inglese *Automatic Speech Recognition*), che coinvolge un gran numero di discipline, quali l'ormai consolidata teoria dei segnali, la fonetica, l'informatica, la psicologia e l'intelligenza artificiale.

Le applicazioni dell'ASR sono innumerevoli. Per fare un solo esempio, si pensi a un grande istituto bancario o a una società assicuratrice, le cui filiali hanno l'esigenza di scambiare quotidianamente informazioni con il centro di calcolo. Attualmente lo scambio di dati avviene mediante terminali video o telescriventi, apparecchiature relativamente costose. Inoltre l'operatore deve essere a conoscenza del linguaggio di accesso alla base di dati. Utilizzando un'apparecchiatura ASR, ciascuna filiale potrebbe comunicare con il calcolatore centrale mediante un semplice telefono parlando in linguaggio naturale, anche se con alcuni vincoli lessicali e sintattici.

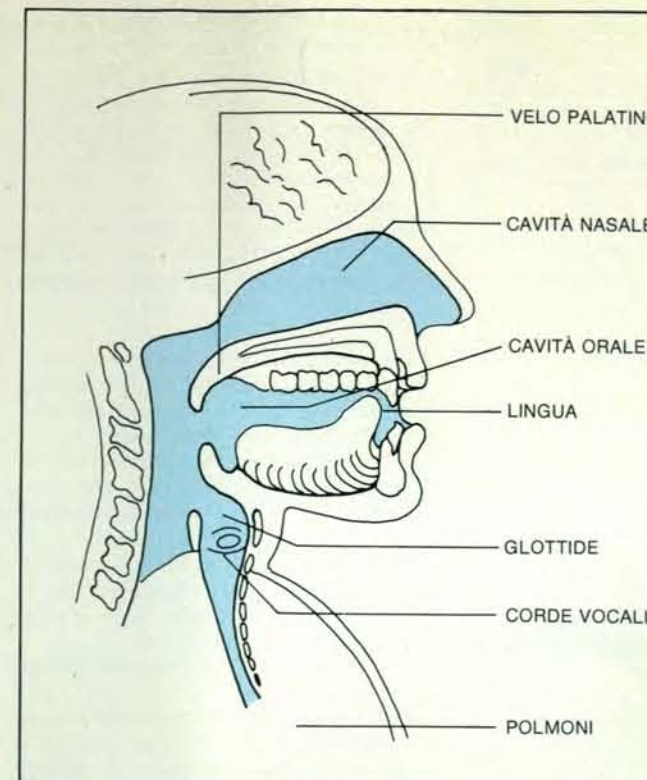
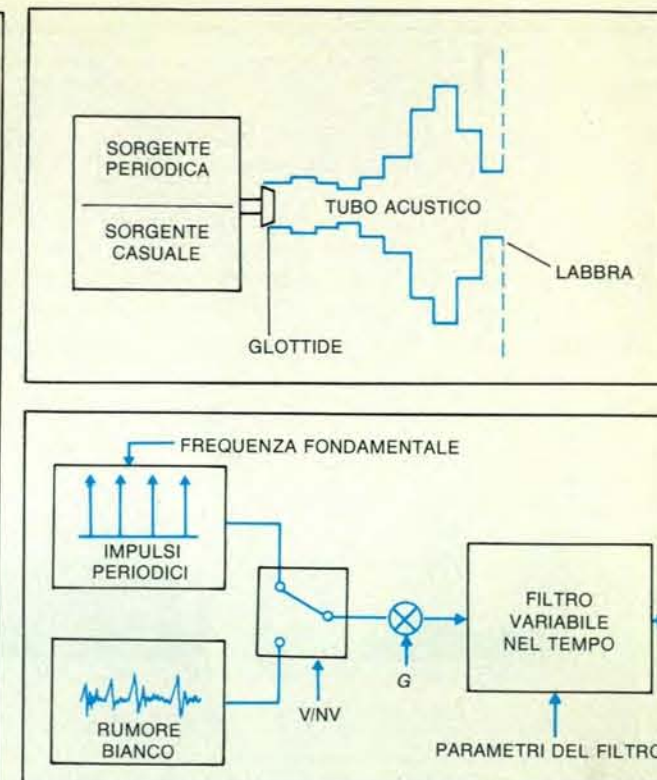
Siamo ancora lontani da applicazioni pratiche di questo genere, ma non siamo più nemmeno nel campo della fantascienza: già oggi siamo in grado di costruire prototipi da laboratorio capaci di comprendere brevi frasi con una sintassi relativamente rigida e appartenenti a un dominio semantico ben delimitato.

L'apparato vocale può essere schematizzato come un tubo acustico in grado di variare la propria forma nel tempo, eccitato da una sorgente di energia. Il tubo acustico è costituito da quel tratto dell'apparato respiratorio che va dalla glottide alle labbra, eventualmente accoppiato alla cavità nasale, e ha in genere una funzione filtrante sul segnale alimentato dalla sorgente di energia, che può essere individuata nei polmoni.

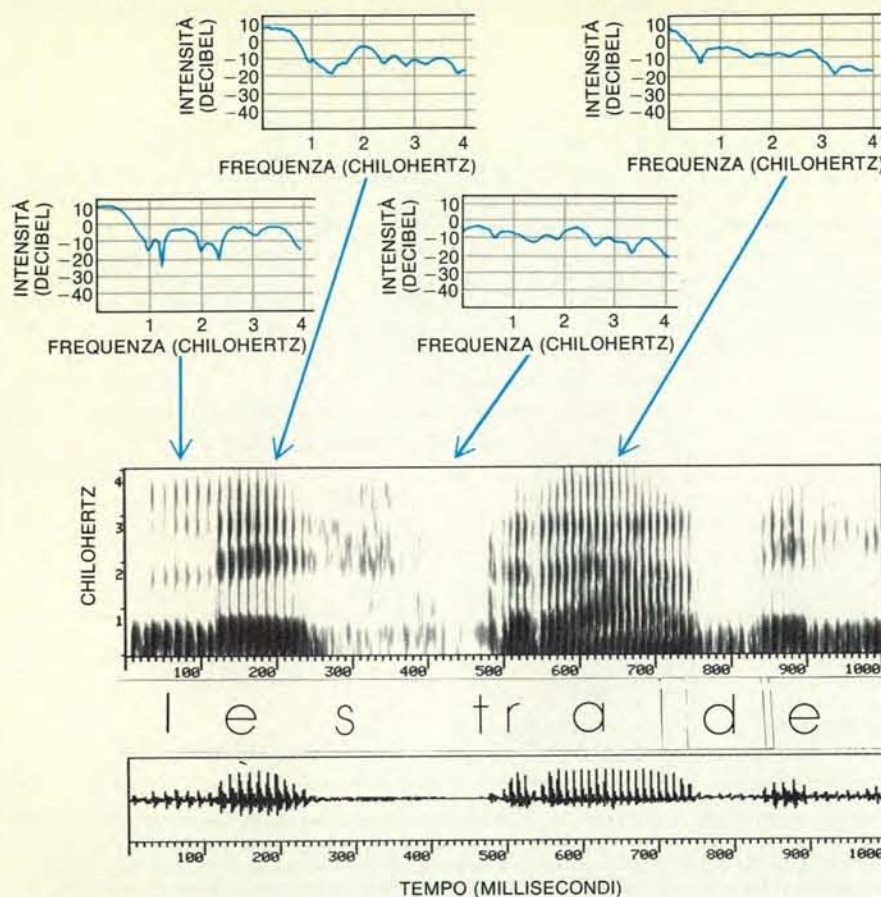
È possibile altresì individuare diversi meccanismi mediante i quali siamo in grado di produrre tutti i suoni di una lingua (faremo riferimento nel seguito alla lingua italiana). Il primo, responsabile dei suoni cosiddetti vocalizzati, utilizza un'apertura, le «corde vocali», che si trova nella laringe all'altezza della glottide. Tale apertura è in grado di aprirsi e chiudersi a un ritmo (detto frequenza fondamentale) variabile dagli 80 ai 200 periodi al secondo (nel caso del parlato normale), producendo così un segnale acustico periodico di forma quasi triangolare. Il segnale, passando attraverso il tubo acustico, viene filtrato acquistando caratteristiche spettrali determinate dalla forma assunta dal tubo acustico stesso.

Le vocali sono suoni (o fonemi) vocalizzati. A esempio nella /i/ il tubo acustico assume una configurazione nella quale si può individuare un punto di costrizione molto vicino alle labbra, mentre nel pronunciare una /o/ la costrizione si trova più indietro, verso la glottide. Esistono però anche alcune consonanti che possono essere considerate vocalizzate. Quando vengono pronunciate una /m/ o una /n/ (consonanti nasali), utilizziamo lo stesso meccanismo delle vocali; l'unica differenza sta nel fatto che in questo caso il velo palatino è abbassato, permettendo all'aria proveniente dalle corde vocali di fluire attraverso la cavità nasale, che si comporta quindi come un secondo tubo acustico posto in parallelo a quello principale.

Le fricative sorde (/f/, /s/, /c/ dolce come nella parola *cena*) sono prodotte invece mediante un diverso meccanismo. Il segnale filtrato dal tubo acustico non è più quello prodotto dalle corde vocali, bensì viene generato dalla turbolenza del flusso d'aria che si crea in corrispondenza di una costrizione del cavo orale. Questo segnale ha caratteristiche di rumore e quindi non presenta una struttura periodica. La pronuncia della /g/ dolce (come in *gelo*) della /v/ e della /z/ (fricative sonore) utilizza entrambe le sorgenti viste precedentemente; infatti il tubo acustico è eccitato da un segnale prodotto in parte dalle corde vocali e in parte dalla turbolenza dell'aria.

Esiste infine un'altra categoria di suoni prodotta con un meccanismo totalmente diverso dai precedenti. Tali suoni sono detti esplosivi o occlusivi e comprendono i fonemi /p/, /t/, /c/ dura come in *cane*, /b/, /d/, /g/ dura come in *ghiaccio*. Durante la pronuncia di questi fonemi, l'aria viene compressa nella cavità orale mediante la chiusura delle labbra per un tempo dell'ordine dei 100 millisecondi. Alla fine di questa fase, detta stop, la seconda fase, consistente in una veloce fuoriuscita dell'aria (esplosione), produce un breve impulso acustico di considerevole intensità. Anche l'esplosione può essere vocalizzata (come nelle esplosive sonore /b/, /d/, /g/ di *ghiaccio*) se durante la fuoriuscita d'aria vengono utilizzate le corde vocali.

La sezione sagittale dell'apparato fonatorio (*schema a sinistra*) può essere rappresentata come un tubo acustico cilindrico a sezione variabile eccitato da due diverse sorgenti (*in alto a destra*); la forma del tubo varia nel tempo con l'evolversi del fenomeno acustico. La sorgente periodica (individuabile nelle corde vocali, le quali aprendosi e chiudendosi ritmicamente trasformano il flusso d'aria proveniente dai polmoni in un segnale periodico) è responsabile dei suoni vocalizzati (per esempio vocali, nasali e liquide). Un altro tipo di eccitazione è dovuto alla sorgente casuale (la quale è causata dalla turbolenza dell'aria che si crea in prossimità di una occasionale ostruzione della cavità orale e genera un segnale le cui caratteristiche sono quelle del rumore bianco); essa è responsabile dei suoni non vocalizzati (fricative e occlusive sorde, per esempio). In basso a destra è riportato un modello elettrico semplificato del meccanismo di produzione della voce. Attraverso il commutatore V/NV (vocalizzato/non vocalizzato) viene selezionata la sorgente di eccitazione; il segnale è amplificato di un fattore *G* e filtrato attraverso un filtro elettrico le cui caratteristiche variano nel tempo come le caratteristiche filtranti del tubo acustico.

I suoni elementari o fonemi (la lingua italiana ne conta circa 30) si contraddistinguono dunque dalla forma che assume il tubo acustico durante la loro produzione e dal tipo di eccitazione (vocalizzata, non vocalizzata, esplosione ecc.). Ma esistono altre caratteristiche, dette soprasegmentali, che pur non essendo distintive intervengono durante il normale eloquio conferendo alla lingua parlata quella naturalezza e quella ricchezza di informazioni sulla sintassi, sul significato delle frasi, sullo stato d'animo di chi parla, che la contraddistinguono da tutti gli altri mezzi di comunicazione. Queste caratteristiche sono essenzialmente: la durata dei singoli suoni, la loro intensità e la frequenza fondamentale durante la vocalizzazione. Sappiamo per esempio che al fondo di una frase interrogativa si verifica un aumento della frequenza di vibrazione delle corde vocali, come una diminuzione è osservabile nei suoni vocalizzati in finale di una frase esclamativa. È anche noto, per esempio, che si ha un accorciamento rispetto alla normale durata di una vocale in finale di parola quando questa è seguita da una parola iniziante per vocale (tale accostamento è detto sinalefe), oppure che l'intensità dell'onda sonora aumenta in corrispondenza del soggetto grammaticale di una frase dichiarativa. Questi fenomeni e moltissimi altri che contribuiscono alla prosodia della lingua parlata sono molto difficili da analizzare e da quantificare a causa della loro marcata dipendenza da tutte quelle sorgenti di conoscenza (fonetica, lessicale, sintattica, semantica, pragmatica) che intervengono nella produzione della voce e sono attualmente oggetto di studio, specialmente da parte dei linguisti e degli psicologi sperimentali.

Il fonema è un'astrazione teorica: non esiste come entità a sé stante. Infatti le parole sono formate da sequenze di suoni e le frasi da sequenze di parole; quindi, nel passaggio da un suono all'altro il tubo acustico passa da una configurazione alla successiva in modo continuo, producendo intervalli di segnale le cui caratteristiche spettrali variano in modo molto rapido (transizioni). Perciò i fonemi nel parlato corrente si realizzano in un transitorio iniziale, nel quale le caratteristiche spettrali del segnale si evolvono da quelle del fonema precedente a quelle del fonema in questione, in una zona con caratteristiche stazionarie e in un transitorio di evoluzione spettrale verso il fonema successivo. Almeno nelle sue porzioni transitorie, la realizzazione fisica del fonema (detta in genere allofono) è fortemente dipendente dal contesto fonetico in cui è inserita. Ad esempio la /m/ nella parola *amo* si realizza in un segnale acustico differente dalla /m/ della parola *uomini*; per questo si parla di due diversi allofoni del fonema /m/.

Il fenomeno, che prende il nome di coarticolazione, aumenta in modo notevole il numero effettivo di eventi acustici distinti di una lingua parlata. Si possono comunque definire eventi acustici elementari che risultano abbastanza indipendenti dal contesto. Questi elementi sono chiamati difoni e sono definiti come segmenti di segnale acustico che vanno dalla metà della parte stazionaria di un fonema fino alla metà della parte stazionaria del fonema successivo. Perciò i difoni comprendono interamente

Il sonogramma è una rappresentazione della voce su tre assi; l'asse orizzontale rappresenta il tempo (in millisecondi), l'asse verticale la frequenza (in chilohertz) mentre l'asse perpendicolare al foglio, rappresentato dal maggiore o minore annerimento delle linee, corrisponde al contributo energetico per una data frequenza a un dato istante. Si notino le striature orizzontali, dette formanti, caratteristiche dei suoni vocalizzati. Le formanti corrispondono alle frequenze di risonanza del tratto vocale, quando esso assume una determinata configurazione articolatoria. In figura esse sono ben visibili in corrispondenza dei suoni / l /, / e /, / r /, / a /. Osservando la forma d'onda riportata sotto il sonogramma si può notare la marcata periodicità dei suoni vocalizzati, mentre la fricativa / s / denota caratteristiche di rumore. In corrispondenza delle esplosive / t / e / d / è visibile la quasi totale assenza di energia precedente all'esplosione. Nei riquadri in alto sono riportati gli spettri dei segmenti di segnale indicati dalle frecce.

la transizione fra due fonemi e, se per esempio una lingua prevede 30 fonemi, si possono avere 30 × 30 = 900 difoni. Chiaramente la struttura fonologica di una lingua non permette tutti i possibili accostamenti tra i fonemi (ad esempio in italiano non esistono parole con la sequenza /bv/ oppure /vs/) e quindi il numero di difoni si riduce di qualche centinaio.

Un'altra sorgente di variabilità del segnale vocale che produce notevoli implicazioni nel progetto di un sistema di riconoscimento è dovuta alle differenze di pronuncia che si verificano sia fra parlanti differenti, sia nelle frasi e parole pronunciate in tempi diversi dalla stessa persona. Esistono differenze macroscopiche fra i parlanti, essenzialmente dovute al dialetto e ai difetti di pronuncia, che comunque possono essere previste nel progetto di un riconoscitore del parlato; ciò che invece non può essere previsto è la variabilità spettrale nella voce prodotta da diverse persone. Infatti le dimensioni e le caratteristiche fisiologiche dell'apparato vocale variano da persona a persona, come varia il modo di articolare i diversi suoni. Inoltre anche la voce prodotta da uno stesso soggetto rivela una notevole varianza nelle caratteristiche di un dato suono pronunciato in tempi diversi.

La voce deve quindi essere considerata un fenomeno casuale (in termini più esatti, un processo stocastico) e il suo trattamento richiede l'uso di tecniche di tipo statistico le cui prestazioni, ovviamente, non sono prevedibili in modo deterministico ma vanno interpretate secondo il concetto di probabilità.

Un modo efficiente per visualizzare l'evoluzione spettrale del segnale vocale è costituito dal sonogramma. Il sonogramma è una rappresentazione su tre assi; tipicamente l'asse orizzontale rappresenta il tempo, quello verticale la frequenza e il terzo, visualizzato dal maggiore o minore annerimento del disegno, l'intensità. È così possibile seguire l'andamento dei contributi energetici alle varie frequenze durante la pronuncia di una data frase; in particolare è molto significativo l'andamento delle tipiche striature orizzontali (chiamate formanti) che corrispondono alle frequenze di risonanza del tratto vocale e sono quindi direttamente correlate all'evoluzione della sua configurazione articolatoria.

Le tecniche per riconoscere automaticamente la voce sono notevolmente complesse, e non possono essere realizzate mediante sistemi analogici, ma richiedono l'uso di un calcolatore numerico come unità di elaborazione. Ma come è possibile introdurre la voce in un calcolatore? Il problema viene risolto trasformando il segnale elettrico fornito da un microfono in una successione di numeri in codice binario, direttamente utilizzabili da un calcolatore.

La possibilità di effettuare questa numerizzazione dei segnali ci viene garantita dal teorema del campionamento, formulato da H. Nyquist negli anni venti. Questo teorema afferma che un segnale continuo può essere completamente rappresentato e perfettamente ricostruito attraverso una serie di misure, o campioni, effettuate sulla sua ampiezza a regolari intervalli di tempo. L'intervallo fra tali campioni non deve però essere superiore al semiperiodo della più alta frequenza presente nel segnale stesso. Quindi se il segnale non contiene frequenze superiori a 4000 hertz (e questo può essere garantito da un filtraggio passa-basso effettuato sul segnale prima del campionamento) i campioni devono essere estratti con un ritmo di almeno 8000 al secondo (si veda anche l'articolo *La riproduzione digitale del suono* di John Monforte, in «Le Scienze», n. 198, febbraio 1985). Questa rappresentazione è però ancora troppo complessa per poter essere efficientemente elaborata.

Occorre inoltre considerare che l'informazione contenuta nella forma d'onda del segnale vocale è affetta da una notevole ridondanza. Cioè, gran parte di tale informazione può essere eliminata mantenendo inalterate le caratteristiche del segnale che rendono i vari suoni percettivamente diversi. Dobbiamo quindi riuscire a trasformare la forma d'onda in una configurazione più semplice che contenga possibilmente tutta e sola l'informazione discriminante i vari eventi fonetici. Una rappresentazione di questo tipo viene in genere indicata con il termine *pattern* (configurazione).

Esistono diversi metodi per estrarre una configurazione del segnale vocale e ancora non si conosce con certezza quale di essi permetta di ottenere in assoluto le migliori prestazioni al fine del riconoscimento della voce. Uno di questi è il cosiddetto metodo delle bande critiche.

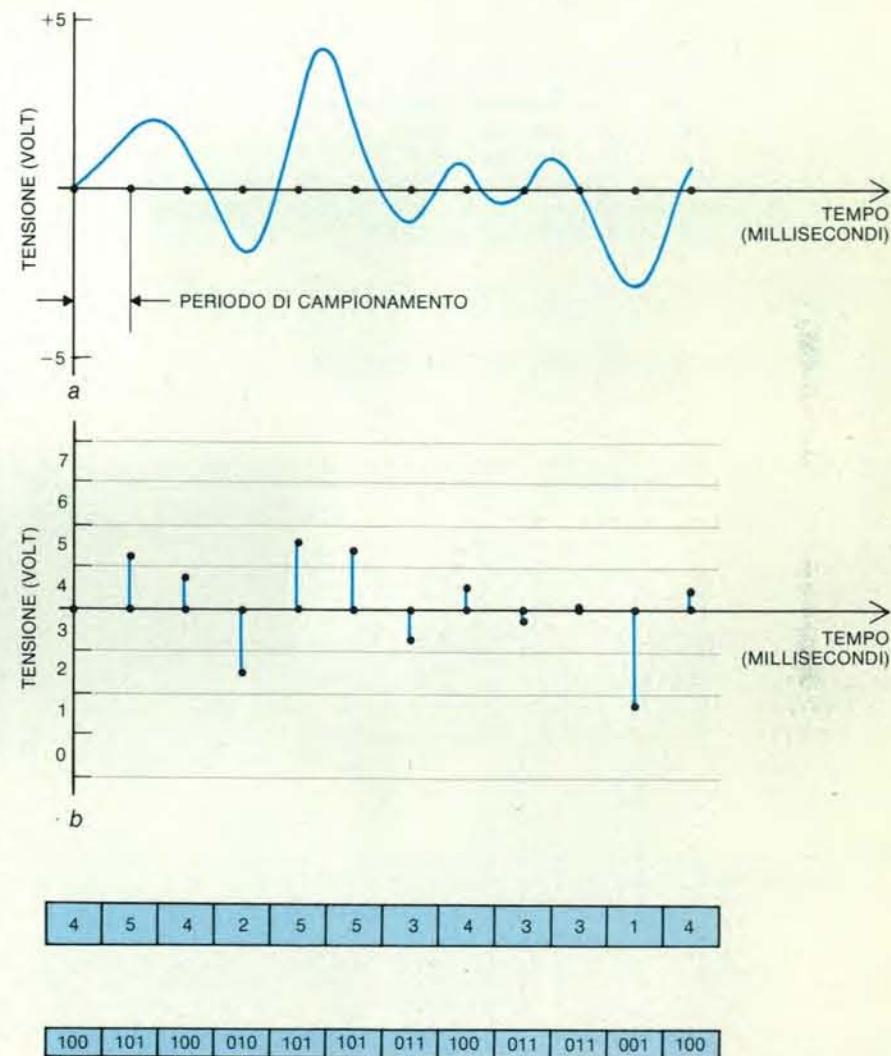
L'orecchio umano è capace di sintonizzarsi su un certo numero di bande di frequenza (bande critiche o articolatorie), rilevabili con esperimenti di tipo percettivo. È ragionevole pensare che questo meccanismo, probabilmente alla base della percezione umana, opportunamente simulato dia buoni risultati nel riconoscimento automatico del parlato.

Abbiamo visto che il segnale vocale è in continua evoluzione spettrale; ciononostante, date le caratteristiche meccaniche dell'apparato fonatorio, può, con buona approssimazione, essere ritenuto stazionario (dal punto di vista delle caratteristiche spettrali) in intervalli dell'ordine dei millisecondi. Supponiamo quindi di suddividere il segnale in intervalli consecutivi della durata di 10 millisecondi che chiameremo «finestre» (*frame* in inglese). Ciascuna finestra di segnale possiede ben determinate caratteristiche spettrali, rilevabili calcolandone lo spettro di energia. (Lo spettro di energia è una curva dell'energia in funzione della frequenza. L'area di tale curva fra due valori di frequenza f_1 e f_2 è proporzionale al contributo energetico al segnale dell'intervallo di frequenze f_1-f_2.)

Quindi ogni finestra può essere descritta mediante i vari contributi energetici in ciascuna banda critica. Per fare un esempio, nell'intervallo fra 300 e 3400 hertz (intervallo di frequenze utilizzato normalmente nelle comunicazioni telefoniche) sono individuabili 13 bande critiche (la prima fra 300 e 430, l'ultima fra 2968 e 3400 hertz). Calcolando quindi lo spettro di energia possiamo ricavare il contributo energetico di ciascuna delle 13 bande e avere una descrizione spettrale della finestra in questione sotto forma di una lista di 13 numeri. In questo modo abbiamo ottenuto la configurazione di una singola finestra. Man mano che il segnale si evolve nel tempo, otteniamo così una successione di liste (o vettori) di numeri, una ogni 10 millisecondi. Chiameremo questa successione «rappresentazione parametrica» del segnale vocale.

Se vogliamo costruire un sistema che operi in tempo reale, tutte le operazioni di calcolo dello spettro, del successivo calcolo delle energie nelle bande critiche e di memorizzazione dei risultati devono essere svolte in un tempo inferiore a 10 millisecondi. Un calcolatore tradizionale non specializzato non è in grado di effettuare queste operazioni in un tempo così limitato. Quindi, in genere, nelle simulazioni di laboratorio vengono usati particolari calcolatori, chiamati *array-processor*, che, collegati a un elaboratore tradizionale, svolgono per questo il grande numero di calcoli che servono, ad esempio, a calcolare lo spettro di una finestra di segnale.

Nei laboratori dello CSELT (Centro studi e laboratori telecomunicazioni) di Torino, il gruppo che si occupa di riconoscimento della voce (costituito da Franco Arcella, Maura Colombo, Marco Cravero, Luciano Fissore, Giorgio Micca, Mario Oreglia, Giancarlo Pirani, Federica Raineri e dall'autore) ha realizzato diversi sistemi sperimentali, utilizzando un calcolatore Digital VAX 11/780. Il primo sistema di cui darò una sommaria descrizione è un riconoscitore di parole pronunciate isolatamente, vale a dire interponendo pause di silenzio fra di loro. Questo è un modo molto innaturale di parlare; infatti normalmente non esistono pause fra le parole di una frase pronunciata correttamente. Anzi, il fenomeno della coarticolazione è presente anche fra le porzioni iniziali e finali di parole consecutive. Però, per compiti molto semplici come il comando di macchine, lo smistamento di pacchi postali o elementari richieste di informazioni, quando cioè una o due parole sono sufficienti, il riconoscimento per parole isolate può essere utilizzato efficacemente.

I sistemi di riconoscimento per parole isolate (IWR, dall'inglese *Isolated Word Recognizer*) sono inoltre molto più semplici dal punto di vista realizzativo ri-

Come si può introdurre la voce in un calcolatore? Il segnale elettrico in uscita da un microfono (*a*) viene campionato a intervalli regolari. Immaginiamo per esempio che il nostro segnale possa assumere valori di tensione compresi fra −5 e +5 volt. Suddividiamo quest'intervallo in un numero *N* di intervalli più piccoli, numerati da 1 a *N*. In questo modo abbiamo costruito un quantizzatore che a ogni campione associa un numero intero (compreso fra 1 e *N*) corrispondente all'intervallo di tensione al quale il campione appartiene (*b*). Il campione può quindi essere rappresentato da questo numero. Quanto più alto è *N*, maggiore è la precisione con cui viene rappresentato ciascun campione. Tipicamente vengono utilizzati 4096 intervalli; quindi ciascun campione può essere espresso da un numero binario di 12 cifre (2^{12} = 4096). Un'apparecchiatura che svolge tutte le operazioni che vanno dal campionamento alla rappresentazione numerica in codice binario di ciascun campione viene chiamata convertitore A/D (analogico/digitale). Mediante un convertitore A/D è quindi possibile inviare alla memoria di un calcolatore una rappresentazione numerica del segnale vocale acquisito mediante un microfono; il teorema del campionamento formulato da H. Nyquist ci assicura che durante questa operazione (a parte le piccole inesattezze introdotte dalla quantizzazione) non si ha perdita di informazione, e pertanto possiamo trattare il segnale numerico come se fosse il segnale reale.

spetto a quelli che riconoscono il parlato continuo. Infatti, essendo le parole separate da pause, il loro inizio e la loro fine sono più facilmente individuabili e per di più non è presente la coarticolazione fra le parole stesse.

Proprio la determinazione dell'inizio e della fine di ciascuna parola (*end-point-detection*) è il blocco iniziale dei sistemi IWR. Questa operazione è molto delicata: dalla sua precisione dipendono le prestazioni del sistema. Generalmente la determinazione dell'inizio e della fine si basa su misure di ampiezza o di energia del segnale in successivi segmenti temporali. Se il sistema opera in ambienti silenziosi questa operazione non presenta grossi problemi. I problemi nascono quando il rumore ambientale è abbastanza elevato, tale da non permettere alle misure di ampiezza di distinguere tra i suoni vocali a bassa intensità (come le fricative /s/ e /f/) e le pause prodotte dal parlante. Inoltre il parlante stesso può provocare rumori indesiderati all'inizio e alla fine della pronuncia delle parole, causati ad esempio dalla apertura e chiusura delle labbra, dal respiro, dai movimenti meccanici del microfono o da eventuali colpi di tosse. In tal caso è necessario utilizzare misure che rendano distinguibile la voce dagli altri rumori.

Estratta la porzione di segnale corrispondente alla parola pronunciata, tale porzione viene elaborata al fine di estrarre la rappresentazione parametrica, cioè la sequenza di configurazioni: su questa opera il successivo processo di riconoscimento.

Neppure l'uomo è in grado di riconoscere e classificare un oggetto se in passato non lo ha mai percepito, o almeno non ha mai acquisito una sua descrizione formale, e questo vale anche per i sistemi di riconoscimento della voce. Per poter riconoscere delle parole, il sistema deve possedere in memoria una loro descrizione. Questa descrizione è fornita alla nostra macchina sotto forma di prototipi di parole pronunciate dal potenziale utente.

È quindi necessario aver fissato in precedenza un vocabolario, cioè l'insieme delle parole che la macchina dovrà riconoscere. Fatto questo, il potenziale utente dovrà pronunciare, almeno una volta, tutte le parole del vocabolario (fase di addestramento). Le loro rappresentazioni parametriche verranno quindi memorizzate ed etichettate (per esempio con un codice che ricorda la forma grafica della parola), e andranno a costituire l'insieme dei prototipi.

Ammettiamo che adesso venga pronunciata una parola del vocabolario. Il riconoscimento di tale parola avviene effettuando un confronto fra la sua rappresentazione parametrica e tutti i prototipi memorizzati durante l'addestramento. Come fare questo confronto? Innanzi tutto occorre definire una metrica, cioè una misura che stabilisca la similarità, o equivalentemente la dissimilarità, fra due configurazioni.

Una configurazione corrispondente a una finestra (10 millisecondi) di segnale è una lista di N numeri, ciascuno dei quali ha un ben preciso significato fisico. Esso può quindi essere immaginato come un punto in uno spazio a N dimensioni (spazio delle caratteristiche); le coordinate di tale punto sono proprio gli N numeri che compongono la configurazione. Sotto questo punto di vista è intuitivamente ragionevole definire una misura di dissimilarità fra due configurazioni come la misura della distanza geometrica che intercorre fra i punti corrispondenti nello spazio delle caratteristiche. Sempre intuitivamente potremmo quindi definire la «distanza» fra un prototipo e la rappresentazione parametrica della parola da riconoscere come la somma delle distanze fra le due configurazioni.

È chiaro che, per effettuare questa operazione, è necessario allineare le due rappresentazioni: mettere cioè in corrispondenza ciascuna configurazione del prototipo con ciascuna configurazione della parola incognita; verranno quindi calcolate le distanze fra le configurazioni corrispondenti, ne verrà fatta la somma e il numero ottenuto potrà essere considerato come una misura di dissimilarità fra le due rappresentazioni. Purtroppo l'operazione di allineamento non è così semplice come potrebbe sembrare.

Immaginiamo di voler allineare le rappresentazioni relative a pronunce diverse della stessa parola. Molto probabilmente potremmo osservare che la lunghezza di tali rappresentazioni (cioè il numero di configurazioni) è diversa. Inoltre saremmo portati a dubitare sulla conservazione delle durate relative dei vari suoni all'interno delle due pronunce. Cioè, in pratica, la stessa parola può essere pronunciata con durate diverse dei vari suoni che la compongono.

L'allineamento quindi deve tener conto di questi fenomeni, deve cioè essere in grado di comprimere o espandere temporalmente le due rappresentazioni in modo da mettere in corrispondenza elementi acusticamente simili (a patto che le due rappresentazioni corrispondano alla stessa parola). Il problema può essere impostato in questi termini: fra tutti i modi possibili in cui possiamo allineare le nostre due rappresentazioni (mettere cioè in corrispondenza le finestre di una con quelle dell'altra) scegliamo quello che meglio fa corrispondere configurazioni simili (cioè relative a eventi simili del segnale vocale). Come valutare la bontà dell'allineamento? Ci viene in aiuto la misura di dissimilarità definita precedentemente. L'allineamento migliore è quello che dà luogo alla minore dissimilarità (somma delle distanze fra le configurazioni corrispondenti) fra le due rappresentazioni. Quindi, almeno in linea teorica, per ogni confronto fra la configurazione da riconoscere e un qualsiasi prototipo dovremmo elencare tutti i possibili allineamenti e scegliere quello che dà luogo alla minore distanza cumulativa. In pratica questa costosa operazione viene risolta in modo più economico con l'aiuto di una tecnica detta «allineamento temporale dinamico» introdotta da H. Sakoe e S. Chiba.

Questa tecnica opera su un reticolo di rappresentazione che comprende un numero finito di punti nodali. Ciascun punto mette in corrispondenza una configurazione della parola incognita con una configurazione di un dato prototipo. Un qualunque percorso congiungente i nodi estremi del reticolo costituisce un possibile allineamento delle due rappresentazioni. Ciascun percorso ha associato un costo pari alla somma delle distanze fra le configurazioni messe in corrispondenza fino a quel punto. L'allineamento migliore è quindi quello corrispondente al percorso a costo minimo. Il percorso a costo minimo può essere individuato evitando di calcolare il costo di tutti i possibili percorsi, ma considerando per ciascun nodo del reticolo solamente il percorso parzialmente migliore (cioè con il minore costo) terminante in tale nodo. Tutti gli altri percorsi confluenti nel nodo in questione, cioè quelli con costo maggiore, vengono eliminati. Questa procedura ci assicura che, dopo aver esaminato tutti i nodi, l'unico percorso che rimane all'estremo del reticolo corrisponde al migliore in senso assoluto. Il costo associato a questo percorso (normalizzato eventualmente in funzione della lunghezza del percorso stesso) è ciò che viene definito come la distanza fra le due rappresentazioni allineate. È intuitivamente e teoricamente ragionevole pensare che la distanza fra le due rappresentazioni parametriche della stessa parola sia, con alta probabilità, inferiore alla distanza che si ottiene confrontando le rappresentazioni di parole diverse. Su questa considerazione è basata la decisione finale del sistema IWR. Infatti, una volta calcolata la distanza fra la rappresentazione della parola pronunciata e ciascun prototipo, il sistema decide in favore della parola associata al prototipo cui compete la distanza minore.

Qual è il costo computazionale di questa tecnica? Per allineare la rappresentazione parametrica della parola da riconoscere con quella di un prototipo, l'elaboratore deve calcolare un numero di distanze fra configurazioni pari al numero di punti nodali del reticolo. In media una parola ha una durata di circa 50 finestre (500 millisecondi). Quindi il reticolo è composto da circa 2500 nodi, e 2500 sono le distanze che devono essere calcolate. Se assumiamo che una configurazione sia composta da 13 numeri, il calcolo di una distanza fra due configurazioni richiede 26 operazioni aritmetiche (13 somme e 13 moltiplicazioni). Quindi, ogni confronto richiede circa 650 000 operazioni aritmetiche. Se il vocabolario è composto da 10 parole (per esempio le 10 cifre) e vogliamo che il sistema riconosca le parole in tempo reale, ammettendo un ritardo di riconoscimento dell'ordine dei 400 millisecondi, per ciascuna parola pronunciata è necessario che l'elaboratore sia in grado di effettuare circa

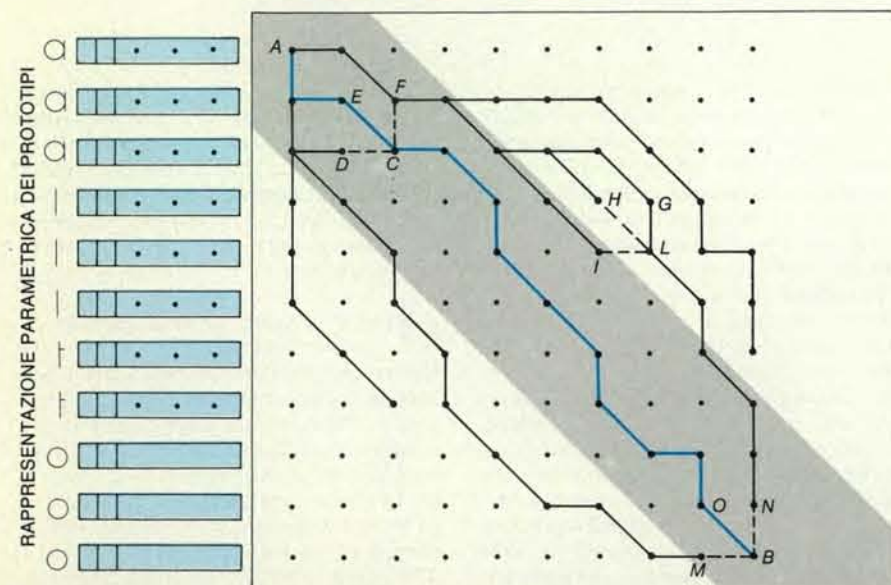
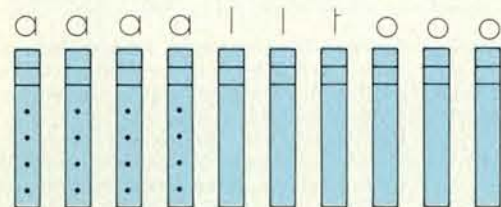
Nel processo di estrazione delle configurazioni, il segnale vocale (*a*) viene suddiviso in finestre, intervalli consecutivi della durata di 10 millisecondi. Di ciascuna finestra si calcola lo spettro di energia (*b*). Poi l'asse delle frequenze viene suddiviso in un determinato numero di bande (bande critiche dell'orecchio). L'area dello spettro di energia in ciascuna banda è proporzionale al contributo energetico di quell'intervallo ed è un elemento della configurazione (*c*). La sequenza delle configurazioni dà la rappresentazione parametrica della parola o della frase.

I sistemi di riconoscimento del parlato più semplici, come quello raffigurato qui, sono in grado di riconoscere parole pronunciate isolatamente. Il segnale vocale viene elaborato in modo da determinare l'inizio e la fine della parola pronunciata. Il segmento di segnale corrispondente viene poi trasformato nella sua rappresentazione parametrica mediante il calcolo delle configurazioni. Durante la fase di addestramento, ciascuna parola del vocabolario (precedentemente fissato) viene pronunciata dal potenziale utente e la sua rappresentazione parametrica viene memorizzata come prototipo. In fase di riconoscimento la rappresentazione parametrica della parola incognita viene allineata con ciascun prototipo e viene calcolata la distanza. Sulla base delle distanze ottenute, il blocco decisionale emette l'ipotesi sulla parola riconosciuta. Normalmente la decisione viene presa in favore della parola il cui prototipo ha dato luogo alla distanza minore.

16 250 000 operazioni aritmetiche al secondo! È evidente che questo è un numero eccessivamente grande anche per un moderno calcolatore. Occorre ridurre di qualche ordine di grandezza il numero di operazioni al fine di poter realizzare un sistema funzionante in tempo reale. A tale scopo viene introdotta una tecnica empirica che limita lo spazio di ricerca del percorso ottimo a una striscia collocata intorno alla diagonale del reticolo di allineamento temporale. È infatti molto verosimile che la distorsione temporale osservabile in pronunce diverse della stessa parola non sia elevata; è quindi probabile che il percorso ottimo di allineamento non risulti eccessivamente deviato dalla diagonale del reticolo. In questo modo, riducendo il reticolo di ricerca a una fascia la cui larghezza tipica è di 6-8 punti, il calcolo di 300-400 distanze è sufficiente per allineare due rappresentazioni.

Esistono inoltre tecniche che permettono di ridurre il numero di finestre di una rappresentazione. Queste tecniche si basano sul fatto che, in zone di stazionarietà spettrale del segnale vocale, le configurazioni sono molto simili fra di loro. Non si ha quindi perdita di informazione spettrale se si rappresentano le porzioni stazionarie con una sola configurazione (questa tecnica viene chiamata «codifica a finestra variabile»: infatti in questo modo la durata temporale di ciascuna finestra non è più costante). Un modo molto semplice di effettuare questa riduzione della ridondanza delle rappresentazioni parametriche consiste nel confrontare (sempre mediante una misura di similarità) ciascuna configurazione estratta dal segnale vocale in ingresso con le successive in ordine temporale, e nel tenere una sola configurazione in quelle sequenze con alta similarità, scartando le rimanenti.

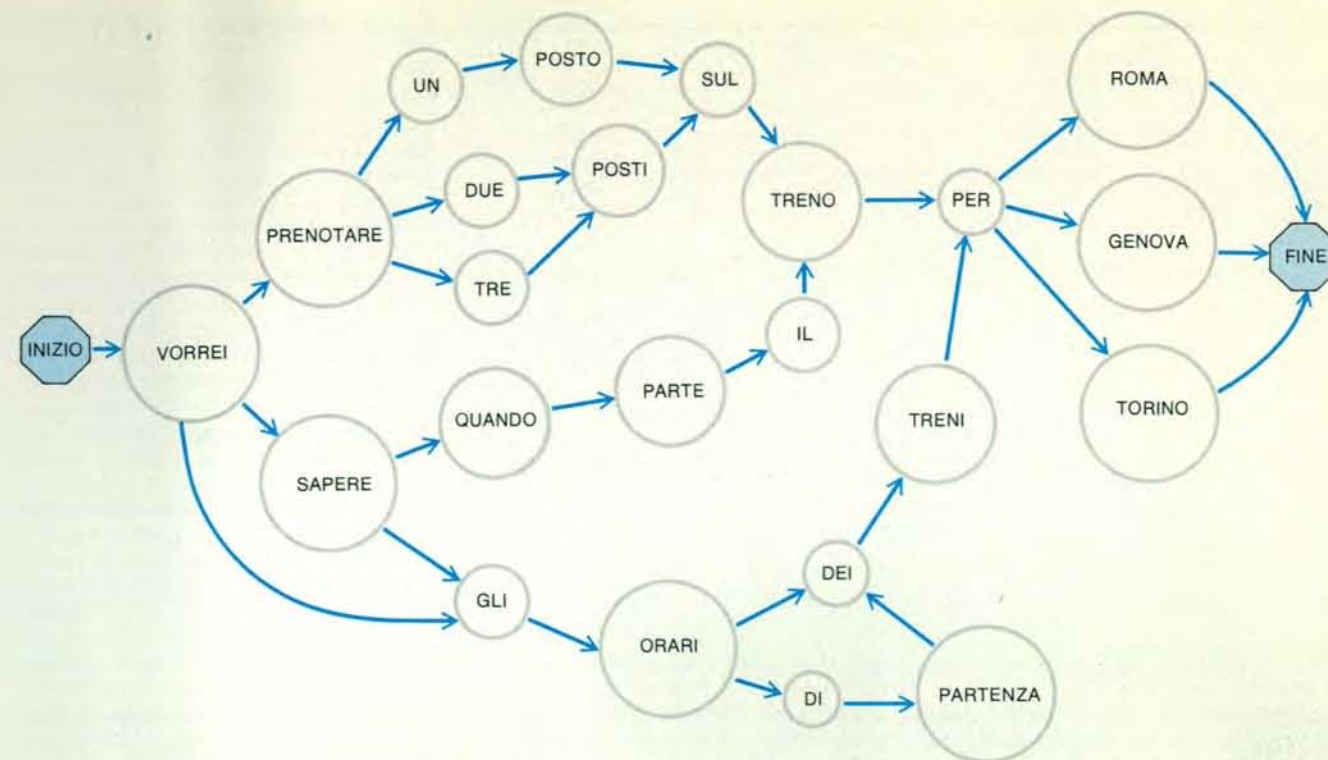
RAPPRESENTAZIONE PARAMETRICA DELLA PAROLA DA RICONOSCERE

RAPPRESENTAZIONE PARAMETRICA DEI PROTOTIPI

Diverse pronunce della stessa parola possono differire notevolmente in durata; inoltre possono essere diverse anche le durate relative dei vari suoni. Quindi, per effettuare un confronto fra due rappresentazioni parametriche (in questo caso fra due repliche della parola «alto») è necessario allinearle temporalmente. Un qualunque percorso congiungente i punti *A* e *B* in figura è un possibile allineamento temporale. Il percorso ottimo è quello per cui è minima la somma delle distanze fra le configurazioni messe in corrispondenza. Questo percorso può essere individuato mediante un algoritmo, detto programmazione dinamica, che esamina tutti i punti del piano procedendo da *A* verso *B*. Per ogni punto viene quindi considerato solo il miglior percorso parziale che inizia in *A* e termina nel nuovo punto stesso. Per esempio, nel punto *C* confluiscono tre percorsi (*A-D-C*, *A-E-C*, *A-F-C*). Viene scelto *A-E-C* perché lungo di esso la somma delle distanze tra le configurazioni è minore rispetto agli altri casi. Lo stesso accade per i percorsi *A-I-L*, *A-H-L*, *A-G-L*. Giunti al punto *B*, per lo stesso motivo viene scelto il percorso *A-O-B*; la somma delle distanze fra le configurazioni messe in corrispondenza da *A-O-B* fornisce la distanza tra le due parole. Poiché il numero di operazioni necessarie a effettuare l'allineamento temporale è molto elevato, in genere si considerano solamente i percorsi all'interno di una fascia (*in grigio*) sovrapposta alla diagonale del reticolo.

Molte altre tecniche più raffinate che agiscono sia sull'allineamento sia sulla dimensione della rappresentazione sono state sperimentate con successo al fine di ridurre la complessità (numero di operazioni al secondo e quantità di memoria necessaria) dei sistemi di riconoscimento, permettendo la realizzazione di simulazioni su elaboratori generici e di prototipi sperimentali (alcuni di essi già in commercio) funzionanti in tempo reale.

Un sistema in grado di riconoscere il parlato continuo, cioè senza pause fra le parole, deve risolvere problemi ulteriori rispetto a un IWR. Non solo non conosce l'esatta identità delle parole contenute in una frase, ma neppure il loro numero, né l'istante di inizio e di fine di ciascuna di esse. Se fossimo in grado di riconoscere il punto di separazione, nel segnale vocale, fra una parola e la successiva, potremmo applicare la tecnica delle parole isolate sulle varie porzioni di segnale corrispondenti alle singole parole. In realtà non esiste alcuna tecnica che permetta di effettuare questa operazione in modo affidabile, anche perché la separazione fra le parole non è facilmente definibile a causa del fenomeno della coarticolazione.

In quale modo riusciamo a comprendere le parole contenute in una frase pronunciata in maniera naturale? Il processo è ancora in parte sconosciuto, ma si può dire che non solo l'evidenza acustica dei singoli suoni contribuisce alla comprensione di una frase. Infatti esperimenti di tipo percettivo hanno dimostrato che anche in condizioni ottime (assenza di rumore ambientale e distorsione del segnale) non è generalmente possibile determinare con alta affidabilità l'identità fonetica dei singoli suoni usando la sola conoscenza acustica. Alcuni esperimenti effettuati all'inizio degli anni settanta soprattutto presso il Massachusetts Institute of Technology hanno provato che l'affidabilità della determinazione può essere notevolmente incrementata usando la ridondanza fornita dalla conoscenza del vocabolario, della sintassi e della semantica. Un sistema evoluto di riconoscimento automatico della voce non deve trascurare questi aspetti.

Verso la metà degli anni settanta le più importanti istituzioni americane operanti nel settore dei calcolatori elettronici (IBM, Carnegie Mellon University, Stanford Research Institute, Bolt Beranek and Newman ecc.) si sono impegnate in un ampio progetto, il progetto ARPA-SUR (Advanced Research Project Agency - Speech Understanding Research), uno dei cui obiettivi era costruire un sistema in grado di comprendere il linguaggio parlato. Sono nati da questo progetto diversi sistemi che hanno sperimentato per primi le tecniche sulle quali si basano praticamente tutti i sistemi attuali. Il progetto ARPA era molto ambizioso e, anche se non ha raggiunto completamente il suo

Una grammatica regolare (ovvero una grammatica di tipo 3), espressa con un automa a stati finiti, può rappresentare formalmente un sottoinsieme, limitato ma pur sempre significativo, del linguaggio. I nodi del grafo, rappresentanti le parole del vocabolario, sono collegati attraverso archi orientati. Qualsiasi percorso dal nodo iniziale al nodo finale genera una frase del linguaggio specificato dalla grammatica.

obiettivo, ha posto solide basi teoriche al riconoscimento automatico del parlato e ha fornito inoltre importanti risultati, in campi come l'intelligenza artificiale e la scienza dei calcolatori, utilizzabili anche al di fuori dell'ASR. Dopo il progetto ARPA si è ad esempio capito che, se è importante avere modelli delle conoscenze ad alto livello (lessicale, sintattica e semantica), ancor più importante è che queste conoscenze siano sostenute da un buon livello acustico. Il sistema HARPY (Carnegie Mellon University, 1976), per esempio, aveva ottime prestazioni, pur con un livello acustico non molto sofisticato, perché forti vincoli di natura sintattica e semantica erano imposti alla struttura delle frasi riconoscibili, sotto forma di grammatica.

Nella terminologia della scienza dell'informazione una grammatica è un insieme di regole che specifica il formalismo di un determinato linguaggio espresso mediante un certo alfabeto.

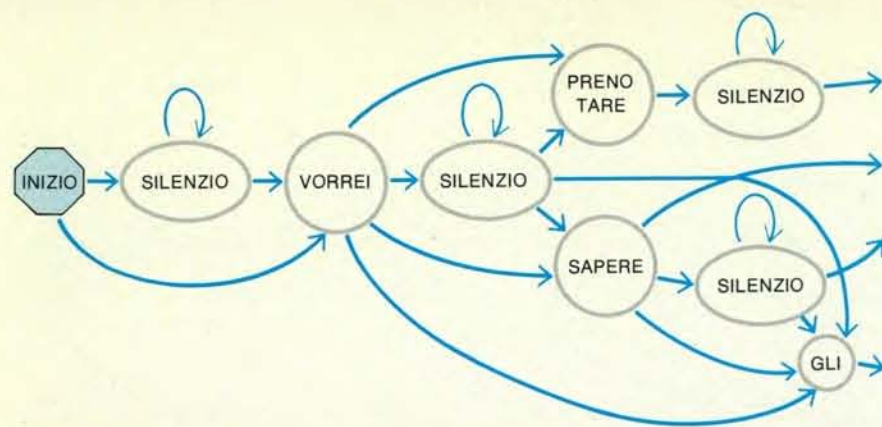
Per esempio, nel linguaggio delle espressioni algebriche l'alfabeto è costituito dalle dieci cifre da 0 a 9, dagli operatori + − × /, dal punto decimale e dai tre tipi di parentesi aperte e chiuse {[()]}. Una qualsiasi successione di tali simboli costituisce una frase, ma non tutte le frasi appartengono al linguaggio delle espressioni algebriche (ad esempio .9()}88+−] non vi appartiene mentre vi appartiene [9+2 × (8.3/5)]−3). Un modo banale di specificare un linguaggio potrebbe essere quello di elencare tutte le frasi possibili, ma molto spesso le frasi sono in numero enorme, se non infinito come nel nostro esempio. Si ricorre quindi a regole che specificano la costruzione delle frasi.

Esistono tecniche per specificare regole di questo tipo in termini formali. Una di queste è costituita dagli automi a stati finiti.

Secondo la forma delle regole che la specificano, una grammatica può appartenere a uno di quattro tipi fondamentali: il tipo 0 corrisponde alle più libere mentre il tipo 3 alle più vincolate. Gli automi a stati finiti possono esprimere solamente grammatiche di tipo 3 (grammatiche regolari) la cui capacità di rappresentazione è la più bassa (ad esempio con una grammatica regolare non è possibile rappresentare il linguaggio delle espressioni algebriche).

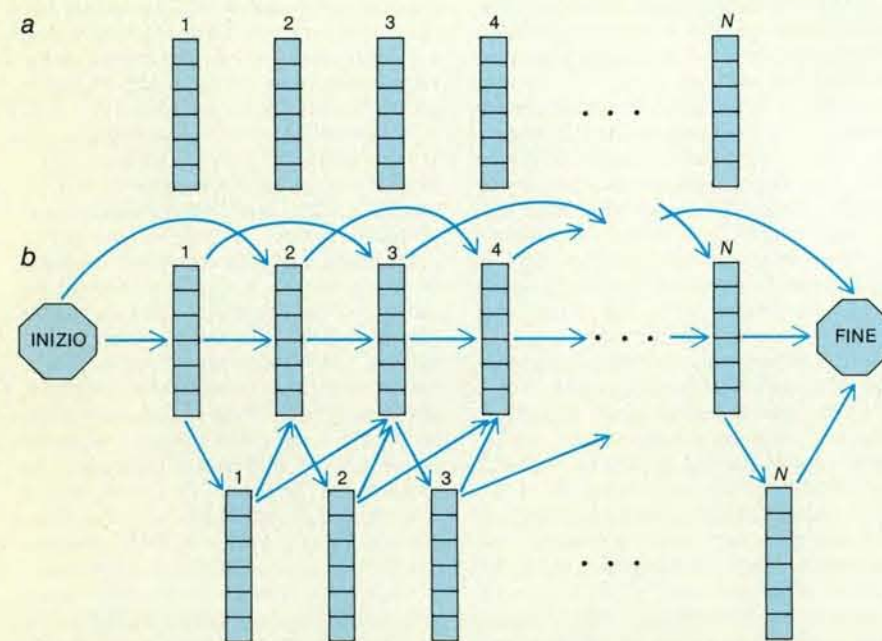
Utilizzando una grammatica regolare è possibile rappresentare un numero elevato di frasi della lingua parlata e quindi specificare un linguaggio nel quale l'alfabeto è costituito dalle parole della lingua. Un automa a stati finiti può essere rappresentato mediante un grafo i cui punti nodali sono collegati mediante archi orientati; se immaginiamo che ciascun nodo rappresenti una parola del linguaggio e che esista un nodo di partenza e uno di arrivo, un qualsiasi percorso congiungente questi ultimi genera una frase sintatticamente congruente con la grammatica specificata dall'automa stesso.

L'introduzione di una grammatica in un sistema di riconoscimento pone un forte vincolo sulle possibili sequenze di parole da riconoscere, evitando quindi di commettere errori che siano incongruenti con la grammatica, ma limita l'insieme di possibili frasi, e quindi costringe l'utente a esprimersi con il formalismo definito (cioè l'utente deve sapere quali sono le sequenze di parole permesse dal linguaggio).

È possibile costruire un sistema di riconoscimento del parlato continuo utilizzando la stessa tecnica vista nel caso delle parole isolate, cioè l'allineamento temporale dinamico, guidato da una conoscenza di tipo sintattico espressa mediante una grammatica regolare. Ammettiamo di avere i prototipi di ciascuna parola del vocabolario e in più un prototipo del rumore ambientale acquisito durante una pausa di silenzio del parlante. Poiché è possibile che durante il colloquio alcune parole siano effettivamente separate da silenzio, è opportuno correggere la grammatica in modo da comprendere nodi di silenzio all'inizio e alla fine della frase e fra le parole; tali nodi possono essere ripetuti indefinitamente (ciò significa non dare nessun vincolo sulla durata del silenzio) o essere saltati. Scendendo a un livello più basso, quello acustico, anche i prototipi possono essere rappresentati mediante automi a stati finiti. Ciascun nodo corrisponde a una finestra di segnale; nel sistema realizzato presso lo CSELT ciascuna finestra può essere ripetuta una

Nella grammatica che rappresenta formalmente un linguaggio si possono introdurre i silenzi: ciascun nodo di silenzio può essere ripetuto indefinitamente oppure saltato. A ciascun nodo di silenzio è associato un prototipo, il quale in genere è costituito da un'unica finestra di silenzio ottenuta durante l'addestramento per tener conto del particolare rumore ambientale.

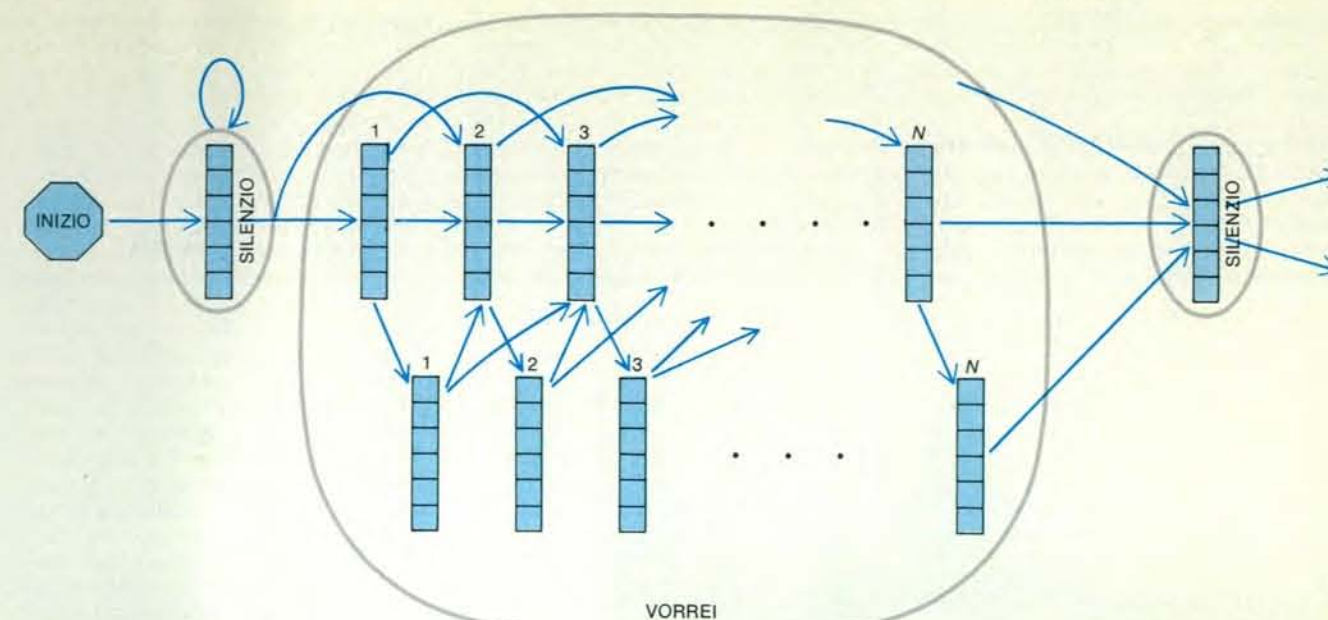
volta o saltata, per ammettere una sorta di allineamento temporale di ingresso.

Sostituendo a ciascun nodo della grammatica di livello più alto (quella che specifica la struttura sintattica delle frasi) la corrispondente grammatica di livello più basso (quella che specifica la struttura spettrale delle parole) si ottiene un automa a stati finiti (rete integrata) in cui a ciascun nodo è assegnata una configurazione di segnale vocale.

A questo punto il problema può essere riportato nei termini visti per le parole isolate. Ci troviamo di fronte a una sequenza di configurazioni, cioè la frase da riconoscere, e dobbiamo trovare quella sequenza nella rete integrata che le è più simile. Utilizzando di nuovo la tecnica di programmazione dinamica siamo in grado di ricavare la sequenza ottima nella rete, quella che dà luogo al valore minimo di distanza cumulata lungo la sequenza stessa. Determinata la sequenza ottima, tale sequenza può essere percorsa a ritroso (fase di *trace-back*) in modo da determinare i nodi della grammatica ad alto livello attraversati, e quindi decodificare le parole contenute nella frase.

La sequenza di configurazioni corrispondente a una parola (*a*) può essere trasformata in un automa a stati finiti (*b*). Per tener conto dei possibili allungamenti o accorciamenti temporali in fase di riconoscimento, ciascuna finestra può essere ripetuta una volta oppure saltata. Si può dimostrare che questa procedura è equivalente all'inserimento di vincoli sui percorsi durante la ricerca del percorso ottimo effettuata con l'allineamento temporale dinamico.

Una prerogativa essenziale di un sistema di riconoscimento in tempo reale è il suo funzionamento in modo continuo. Sarebbe cioè opportuno che il sistema fosse in grado di riconoscere una sequenza continua di parole, senza necessariamente attendere la fine della frase per la decodifica finale. Questo può essere fatto nel sistema CSELT.

Occorre premettere che l'algoritmo di programmazione dinamica procede in sincronismo con le finestre della frase da riconoscere. Per ciascuna finestra di ingresso vengono estese di un passo tutte le sequenze calcolate alla finestra precedente; se il punteggio (o distanza cumulata) di una sequenza è superiore a un determinato valore di soglia, è molto probabile che tale percorso non appartenga alla sequenza ottima e quindi vale la pena di bloccarlo. In questo modo il numero di sequenze da calcolare a ogni nuova finestra rimane praticamente entro limiti accettabili. Può accadere che tutti i percorsi attivi a un certo istante abbiano nel passato un'origine comune, siano cioè ramificazioni di un percorso che, a causa del disattivamento di altri percorsi, è rimasto unico. Quindi, sul tratto in cui tale percorso è unico, le decisioni sulle parole possono essere già prese. Con questa tecnica (detta *partial trace-back*, percorso a ritroso parziale) non solo si può riconoscere una sequenza ininterrotta di parole con un breve ritardo fra l'effettiva pronuncia e il riconoscimento, ma si è in grado di mantenere tollerabili le dimensioni della memoria. Infatti, per permettere il percorso a ritroso, per ogni percorso attivo è necessario memorizzare, oltre ad altre informazioni, la sequenza di nodi da cui è passato; una volta effettuato il percorso a ritroso e decodificato in parole il segmento di frase, tutte le informazioni relative al percorso in questione possono essere cancellate dalla memoria.

Già per un vocabolario di modeste dimensioni (50-100 parole) il numero di operazioni da effettuare nell'unità di tempo è veramente alto. Il progetto *hardware* di una macchina di questo tipo deve senz'altro fare riferimento ad architetture multielaboratore altamente parallelizzate (cioè macchine utilizzanti più calcolatori elementari che svolgono compiti in parallelo). Esistono oggi in commercio alcune macchine progettate per riconoscere il parlato continuo con vocabolari di poche centinaia di parole che sfruttano tecniche del tipo appena descritto, ma i problemi del riconoscimento del parlato continuo non sono completamente risolti: le prestazioni di questi sistemi sono ancora troppo basse, dipendono fortemente dall'ambiente in cui si opera, dalla lingua usata, dall'accuratezza di pronuncia dell'utente; il loro vocabolario è ancora estremamente piccolo per alcune utilizzazioni di interesse, i vincoli sintattici imposti dalla grammatica possono essere troppo pesanti per l'utente. Queste macchine costituiscono comunque un notevole passo in avanti nella scienza dei calcolatori, contribuendo in maniera non trascurabile allo sviluppo e alla ricerca di nuove architetture per i calcolatori della prossima generazione.

Nel caso di un sistema di riconoscimento del parlato continuo, il riconoscimento avviene mediante la ricerca di quel percorso, nella rete integrata, che produce la minore distanza cumulata dalla sequenza di configurazioni della frase da riconoscere. Per generare la rete integrata, a ogni nodo della rete sintattica viene sostituito l'adeguato automa a stati finiti che rappresenta il prototipo della parola corrispondente.

Vi sono molti altri problemi inerenti al riconoscimento di parole isolate o connesse che costituiscono importanti temi di ricerca in tutti i laboratori che si occupano del colloquio vocale uomo-macchina. Uno di questi va sotto il nome di «indipendenza dal parlante». Un sistema di riconoscimento richiede una fase di addestramento la cui lunghezza dipende dalle dimensioni del vocabolario, ma per molte applicazioni sarebbe importante che il sistema non richiedese questa fase. Basti pensare a un servizio pubblico: esso diverrebbe improponibile se ciascun utente dovesse addestrare la macchina su tutto il vocabolario prima di formulare la richiesta. Oppure, per un sistema con vocabolario di qualche migliaio di parole, la fase di addestramento potrebbe diventare una operazione abbastanza onerosa anche se ristretta a un determinato numero di utenti.

Una delle soluzioni proposte fino a oggi è quella fornita dall'analisi dei raggruppamenti *(cluster analysis)*. Riferiamoci a un sistema IWR. Immaginiamo di collezionare prototipi delle parole del vocabolario pronunciate da un'ampia popolazione di parlanti. Tutte le repliche di una determinata parola del vocabolario possono essere confrontate tra loro mediante la tecnica di allineamento dinamico, generando una serie di distanze mutue fra esse. Possiamo immaginare che ciascuna replica sia un punto in uno spazio multidimensionale; la dislocazione di tutti i punti è univocamente determinata dall'insieme di distanze mutue. Se si ammette, e questa è l'ipotesi fondamentale, che esistano evidenti raggruppamenti *(cluster)* di tali punti (cioè repliche della parola in questione che, pur provenendo da parlanti diversi, esibiscono una certa similarità) e se i raggruppamenti sono abbastanza compatti, si può pensare di rappresentare tutte le repliche appartenenti a ciascun raggruppamento mediante un unico prototipo rappresentativo (per esempio il centro di massa del raggruppamento). Il problema fondamentale della determinazione automatica dei raggruppamenti di un insieme di repliche della stessa parola in uno spazio il cui numero di dimensioni è dell'ordine di grandezza del numero di repliche stesse viene in genere risolto mediante tecniche statistiche operanti sugli insiemi di distanze.

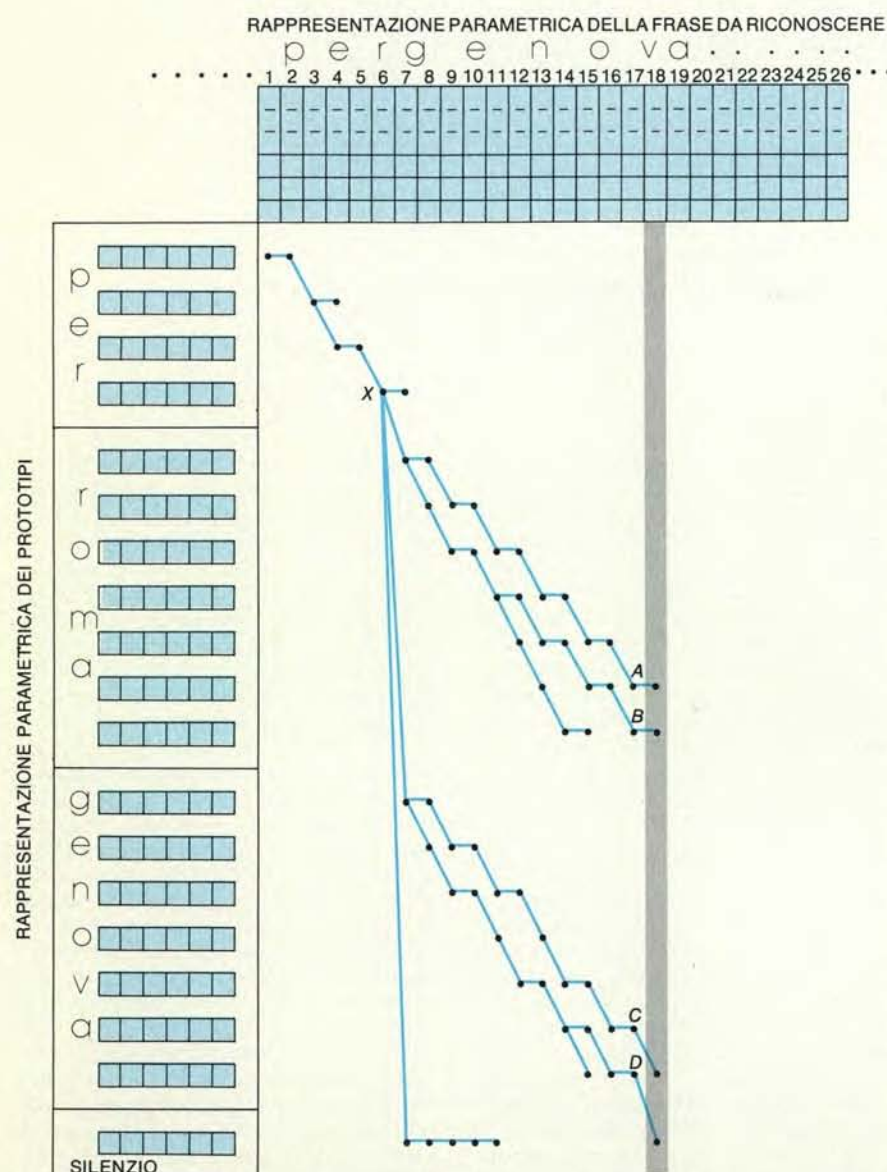
È stato dimostrato che su una popolazione di qualche centinaio di parlanti si evidenziano in media una ventina di raggruppamenti per ciascuna parola. Quindi nel sistema di riconoscimento è opportuno memorizzare più prototipi per ciascuna parola, cioè uno per ciascun raggruppamento.

Questa tecnica non fornisce in generale prestazioni elevate. In genere con un sistema IWR operante con un vocabolario di un centinaio di parole è tipicamente raggiungibile una accuratezza dell'ordine del 99 per cento (99 per cento di parole correttamente riconosciute) se viene effettuato l'addestramento; utilizzando l'analisi dei raggruppamenti le prestazioni scendono al di sotto del 95-96 per cento, senza contare l'enorme aumento di memoria necessaria per i prototipi. Si pensa oggi, come alternativa per alcune applicazioni, di studiare tecniche diverse, che prescindono da un addestramento «universale», ma che cercano di adattare i prototipi al parlante attuale, utilizzando ad esempio misure effettuate su poche frasi preliminari al colloquio vero e proprio.

Le tecniche di riconoscimento che abbiamo visto sono realizzabili finché il vocabolario si mantiene su dimensioni di qualche centinaio di parole. Per una applicazione con un vocabolario abbastanza grande (2000-3000 parole) comincerebbero a sorgere grossi problemi di memorizzazione e di tempo di calcolo; non solo, ma volendo introdurre nuove parole nel vocabolario si renderebbe necessario addestrare di nuovo la macchina. A questo proposito il gruppo di ricerca dello CSELT sta studiando tecniche che utilizzano unità elementari di riconoscimento di dimensioni più piccole della parola. I difoni, almeno per la lingua italiana, sono i candidati più promettenti a rivestire questo ruolo. Con un insieme di circa 400 difoni è possibile costruire la maggior parte delle parole della lingua italiana. L'idea è quella di memorizzare prototipi dei difoni mediante i quali sintetizzare i prototipi delle parole necessarie in ciascuna fase del processo di riconoscimento. I problemi riguardano essenzialmente l'estrazione dei prototipi di difoni. Si può pensare inizialmente di estrarre in modo manuale i prototipi, per esempio osservando su un video grafico la forma d'onda di una

parola e segmentando i difoni che ci interessano, ma questa procedura è molto lunga e richiede personale tecnico con esperienza nella segmentazione di eventi acustici.

Utilizzando modelli più complessi dei prototipi, modelli in grado di acquisire automaticamente conoscenze di tipo statistico sull'evento in questione, è possibile realizzare una procedura completamente automatica di addestramento per difoni partendo da un insieme di parole (non necessariamente appartenenti al vocabolario) che comprenda più casi dei difoni in questione. Questa tecnica, una volta affinata, porterà probabilmente alla realizzazione di sistemi in grado di riconoscere frasi appartenenti a vocabolari di dimensioni dell'ordine del migliaio di parole.

L'impostazione seguita ai laboratori CSELT non è l'unica possibile. Un'altra impostazione significativa è quella che si può definire del «riconoscimento per regole», associata in particolare al nome di V. Zue del Massachusetts Institute of Technology, e basata sul postulato che il segnale vocale contiene tutte le informazioni necessarie al riconoscimento e alla comprensione di una frase. Se l'uomo è in grado di decodificare le informazioni fonetiche contenute in una rappresentazione spettrale del segnale vocale (sonogramma), come Zue ha dimostrato alla fine degli anni settanta, questa abilità può essere espressa mediante regole codificate in un calcolatore. Si tratta quindi di utilizzare le tecniche di intelligenza artificiale conosciute, le quali permettono di inserire un bagaglio di conoscenza in una macchina e di sfruttarlo per formulare e verificare ipotesi sugli stimoli fisici, non secondo uno schema prefissato (e quindi algoritmico), ma utilizzando regole inferenziali che ricalcano le regole utilizzate dall'uomo nel risolvere lo stesso problema. Questa strategia è molto ambiziosa e il suo perseguimento è molto costoso. Purtroppo i risultati non sono immediati; non abbiamo ancora una conoscenza del fenomeno acustico tanto profonda da garantire la realizzazione di una macchina per il riconoscimento della voce con prestazioni paragonabili a quelle dell'uomo. L'acquisizione di questa conoscenza è molto difficile e costosa poiché si basa unicamente sull'osservazione umana di un numero enorme di realizzazioni di eventi acustici. Per questo motivo, il metodo statistico è stato in grado di fornire risultati parziali in breve tempo, che potranno un giorno essere superati da quelli raggiungibili con il metodo fonetico. Probabilmente, però, non esiste fra i due un metodo vincente. La statistica è un mezzo per acquisire conoscenza da grandi quantità di dati, ma è priva di ciò che possiamo chiamare buonsenso. Il buonsenso è l'abilità prettamente umana che permette di non compiere errori grossolani. Molto spesso accade che algoritmi di tipo statistico commettano errori grossolani nella determinazione di parametri della voce, per esempio, tipicamente, nel decidere se un segmento di segnale è vocalizzato o non vocalizzato, laddove un osservatore esperto potrebbe decidere senza dubbio, osservando la forma d'onda o lo spettro di tale segmento. Le tecniche di tipo euristico, come il riconoscimento fonetico, tentano di inserire nei sistemi questa esperienza umana, ma è molto costoso consolidarla mediante un gran numero di osservazioni. Penso che il futuro del riconoscimento automatico della voce (ed è questa la tendenza attuale della ricerca presso lo CSELT) stia nell'unione di queste due filosofie, unione che fornirà ai metodi statistici il buonsenso di cui sono privi e andrà ad arricchire, mediante la statistica, il bagaglio di conoscenze su cui si basano i metodi euristici.

Ammettiamo che un sistema di riconoscimento in tempo reale, dotato della caratteristica di *traceback* (ricerca a ritroso) parziale, stia elaborando la finestra 18 della frase da riconoscere («*...per Genova...*»). I percorsi attivi alla finestra precedente (*A, B, C, D*) vengono estesi secondo le regole definite dalle grammatiche relative ai prototipi e alla sintassi del linguaggio, incrementando il loro costo della distanza fra la configurazione del frame 18 e la configurazione del prototipo raggiunto. Non vengono considerati quei percorsi il cui costo supera di un certo valore (fissato dal progettista) il costo associato al percorso localmente ottimo. Può accadere che tutti i percorsi attivi provengano da un unico nodo (*nodo X nell'esempio*), quindi il percorso da quel nodo a ritroso verso l'inizio è determinato univocamente. I prototipi interessati da questo percorso parziale forniscono le parole riconosciute (*prototipo della parola «per» nell'esempio*). Le informazioni associate al percorso possono quindi essere cancellate, memorizzando unicamente il codice delle parole riconosciute. Questo codice, risultato finale del processo di riconoscimento, può poi essere utilizzato dagli altri livelli di conoscenza del sistema.

I vulcani e le nubi di Venere

I gas di zolfo che formano la coltre nuvolosa di Venere sono forse emessi da vulcani la cui attività sembra confermata sia dalle mappe radar sia dalle analisi chimiche condotte sulla crosta e sull'atmosfera del pianeta

di Ronald G. Prinn

Per un planetologo una delle caratteristiche più interessanti della Terra è la sua attività. Il calore che fluisce dall'interno del pianeta alimenta processi che ne rimodellano continuamente la superficie. Un segno caratteristico di questa attività è il vulcanismo, ossia l'affioramento in superficie di rocce e gas caldi attraverso fessurazioni della crosta. Eruzioni vulcaniche sono state osservate anche su Io, il satellite di Giove, ma non su altri corpi del sistema solare. Eppure Venere, il pianeta a noi più vicino, assomiglia alla Terra sotto molti aspetti: ha circa le stesse dimensioni e la stessa massa, e si è formato nella stessa regione della nebulosa solare in condensazione. I due pianeti dovrebbero quindi aver avuto un'evoluzione simile. Vi sono vulcani attivi su Venere?

Per molto tempo le risposte a questa domanda non potevano essere che ipotetiche, perché uno strato di nubi spesso e persistente ostacolava i tentativi di studiare la superficie venusiana, ma negli ultimi cinque anni la situazione è cambiata. Anche se bloccano la luce visibile, infatti, le nubi sono trasparenti alle onde radio e alle microonde, e ciò ha permesso di realizzare un atlante completo della superficie mediante un radar a bordo di *Pioneer Venus,* in orbita intorno al pianeta fin dal 1978. Le mappe, insieme ad altre immagini radar ad alta risoluzione più recenti ottenute da due sonde spaziali sovietiche e da radiotelescopi a terra, hanno rivelato l'esistenza di strutture di tipo vulcanico.

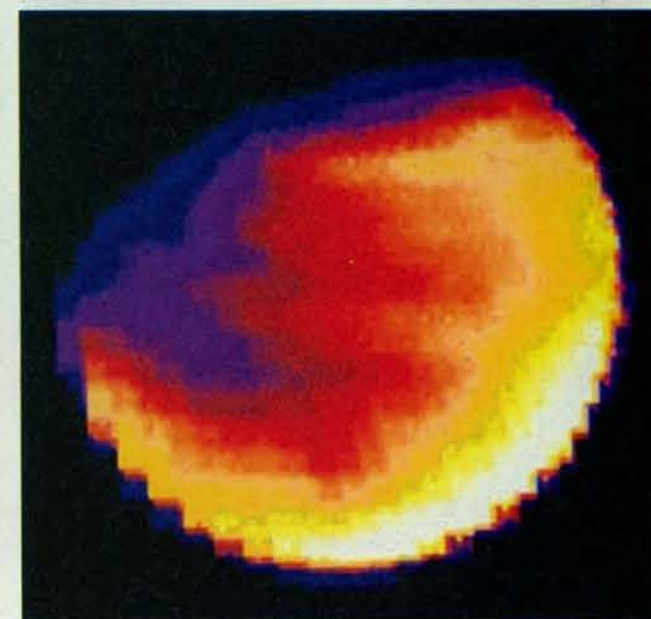
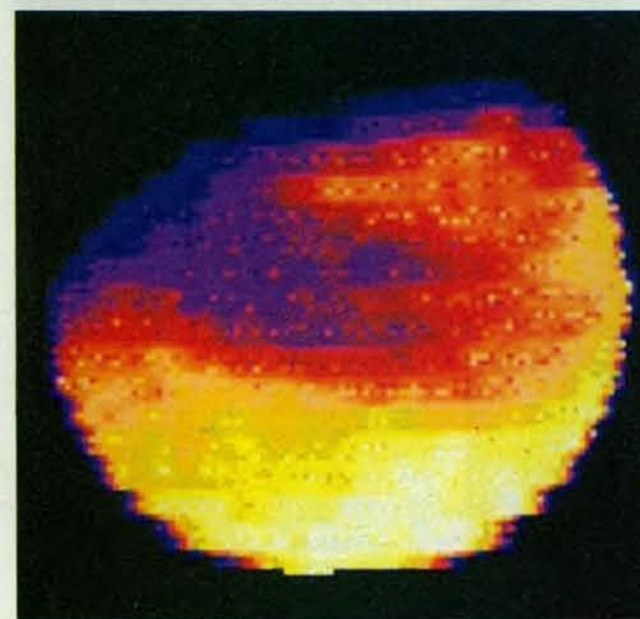
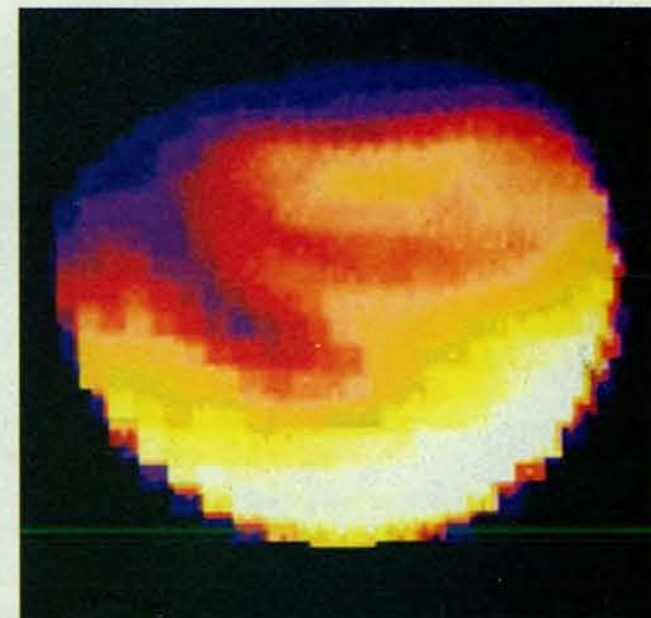
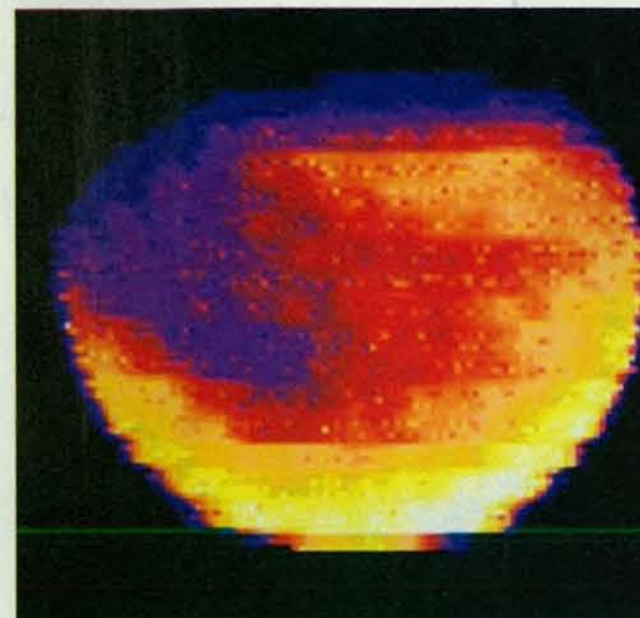
La persistenza stessa delle nubi opache, inoltre, costituisce un'indicazione precisa, anche se indiretta, dell'esistenza di un vulcanismo attivo. Le nubi infatti, che si trovano a quote comprese tra 50 e 70 chilometri, sono composte di acido solforico concentrato e di un materiale che assorbe i raggi ultravioletti, probabilmente zolfo elementare. Negli ultimi anni parecchie sonde hanno resistito all'attacco corrosivo di queste sostanze e all'intenso calore della superficie (460 gradi centigradi) abbastanza a lungo da misurare la composizione dell'atmosfera e della crosta del pianeta. Di conseguenza oggi è possibile interpretare l'interazione di questi due sistemi e comprendere il complesso ciclo di reazioni fotochimiche e termochimiche che trasforma i gas sulfurei nelle particelle delle nubi. Dai dati si ricava che i gas sulfurei vengono iniettati continuamente nell'atmosfera grazie a un meccanismo che non può essere altro che vulcanico. In realtà il rilevamento da parte del modulo orbitale *(orbiter)* di *Pioneer Venus* di variazioni impressionanti nell'abbondanza di anidride carbonica nell'atmosfera oltre la sommità delle nubi fa pensare che Venere sia stato scosso da massicce eruzioni negli ultimi dieci anni.

L'indicazione più diretta dell'esistenza su Venere di vulcani, attivi o estinti, proviene dagli studi radar. Fin dagli anni sessanta sono stati puntati sul pianeta i radiotelescopi di Arecibo a Portorico e di Goldstone in California, i quali hanno fornito la prima valutazione attendibile del raggio (6052 chilometri, contro i 6378 della Terra) e del periodo di rotazione di Venere (243 giorni) oltre alle prime immagini di grandi formazioni superficiali come il grande continente settentrionale di Terra Ishtar. Nel 1977 R. Stephen Saunders e Michael C. Malin del Jet Propulsion Laboratory hanno avanzato l'ipotesi che Mons Theia, nella Regio Beta, sia un grande vulcano a scudo, una struttura approssimativamente simmetrica formata da persistenti colate non esplosive di lava calda che percorre grandi distanze prima di solidificare. Con i suoi 700 chilometri di diametro Mons Theia è molto più grande dei vulcani a scudo delle Hawaii, ma più piccolo del Mons Olympus di Marte che sembra un esempio gigantesco di questo tipo di formazione montuosa.

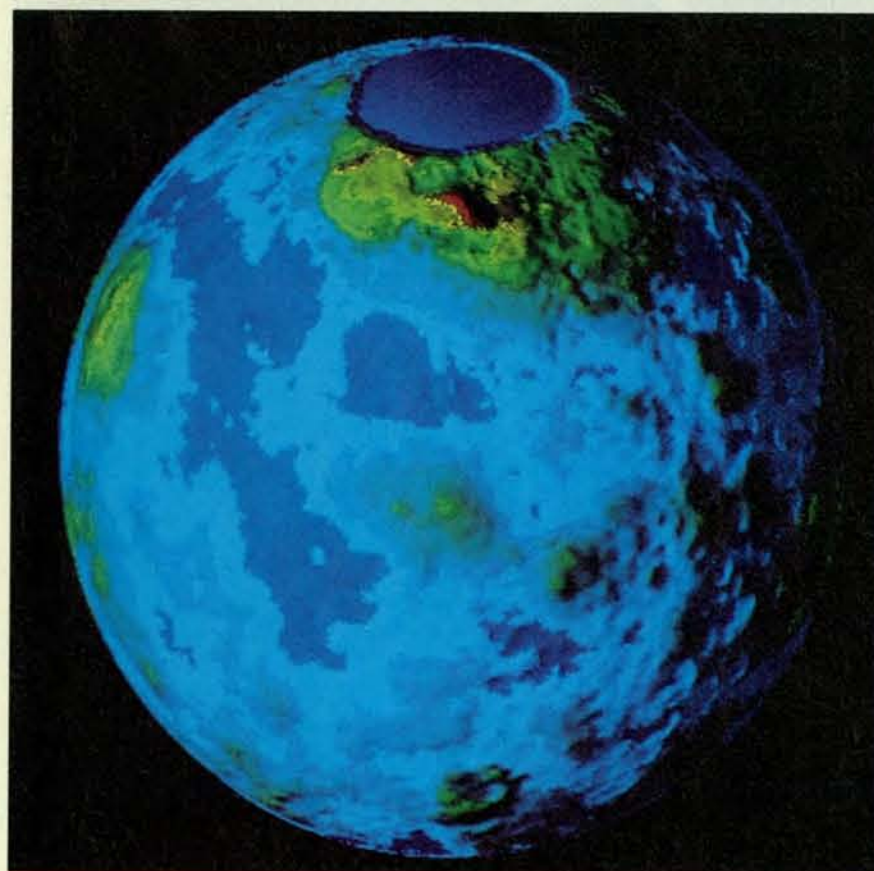
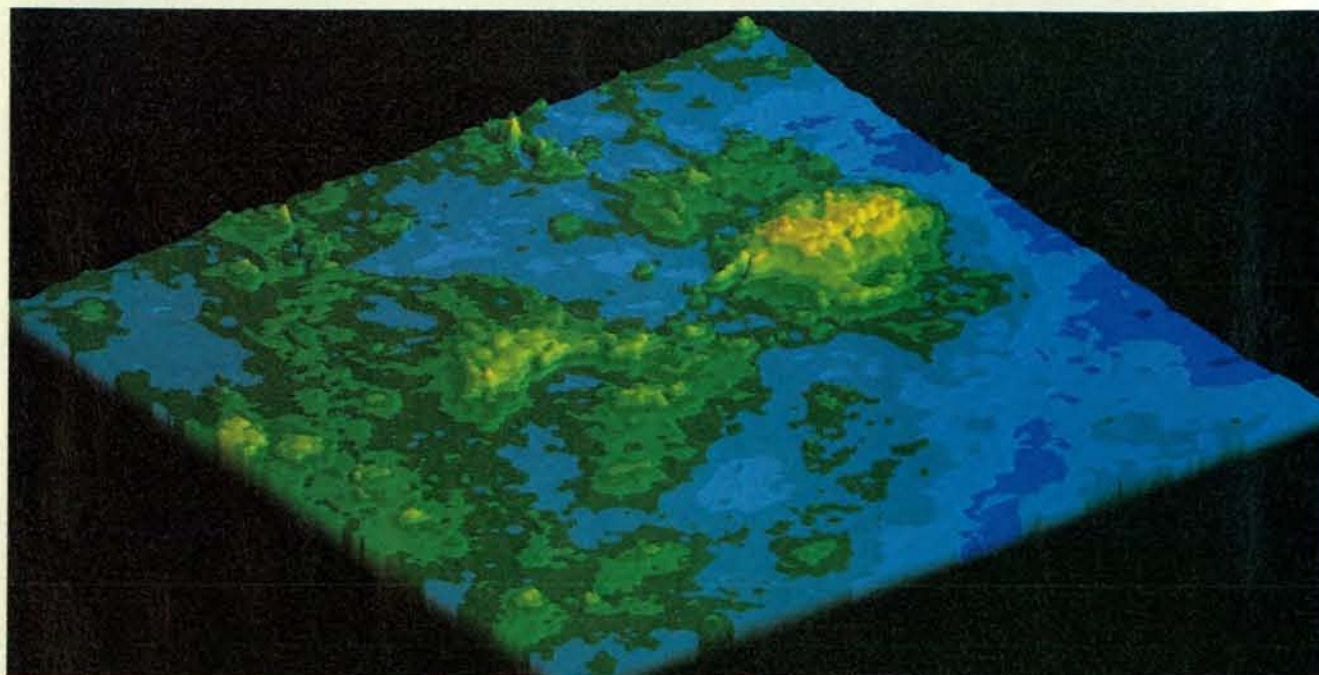
Da quando, nel dicembre 1978, *Pioneer Venus* è entrato in orbita intorno al pianeta, è diventato possibile produrre immagini radar con una risoluzione spaziale superiore e anche cartografare il rilievo superficiale. Nel 1980, sulla base di questi dati Harold Masursky dell'US Geological Survey, Gordon H. Pettengill del Massachusetts Institute of Technology e collaboratori hanno affermato che tutta la Regio Beta, compresi i due rilievi Mons Theia e Mons Rhea (entrambi alti più di 4500 metri), è un'enorme struttura vulcanica formata dal lento accumularsi di lava. In seguito George E. McGill e collaboratori dell'Università del Massachusetts ad Amherst hanno suggerito che i vulcani siano in realtà strutture laviche più modeste sulla sommità di un duomo allungato di crosta sollevata.

Quest'ultima teoria comporta che sotto la Regio Beta abbiano luogo risalite di magma regionali e non locali; l'ipotesi è corroborata dall'osservazione, tratta dalle perturbazioni nell'orbita di *Pioneer Venus,* che nella regione la gravità è più forte della media venusiana. Può darsi che il fenomeno di sollevamento sia abbastanza intenso da causare un certo movimento orizzontale della crosta. Nel 1983 Donald B. Campbell e collaboratori dell'Osservatorio di Arecibo hanno ottenuto un'immagine della Regio Beta con una risoluzione circa 10 volte superiore a quella ottenibile con *Pioneer Venus.* Vi compare una grande frattura *(rift)* rettilinea con a fianco due strutture vulcaniche agli estremi, che taglia l'intera regione. In ogni modo, si concorda, in generale, sul fatto che nella Regio Beta vi sia stato qualche tipo di attività vulcanica. La luminosità delle immagini radar fa pensare che l'attività possa essere stata recente dal punto di vista geologico: è indice infatti di superficie scabra, cioè relativamente giovane e non alterata dagli agenti meteorologici.

La Regio Beta non è l'unica regione in cui siano state osservate strutture vulcaniche. Secondo James W. Head della Brown University e altri ricercatori le vette e le guglie isolate rilevate da *Pioneer Venus* nella Regio Atla, a ovest di Beta vicino all'equatore, possono essere vulcani singoli. Le sonde sovietiche *Venera 15* e *Venera 16,* che si trovano in orbita intorno a Venere dall'ottobre 1983, hanno trasmesso le immagini di grandi formazioni circolari precedentemente non osservate che, avendo diametri di parecchie centinaia di chilometri ed essendo relativamente poco elevate, delimitano la superficie di Terra Ishtar e di altre regioni. A. T. Basilevsky, V. L. Barsukov e collaboratori presso l'Istituto V. I. Vernadsky di geochimica e chimica analitica di Mosca hanno interpretato queste strutture come enormi duomi vulcanici che crollando hanno lasciato pieghe di crosta nel loro intorno.

I radar a bordo dei moduli orbitali delle sonde Venera forniscono immagini con una risoluzione orizzontale compresa tra uno e due chilometri, mentre la risoluzione di *Pioneer Venus* è nel migliore dei casi di 30 chilometri. Purtroppo però le missioni sovietiche sono state progettate per raccogliere immagini di solo un terzo circa del pianeta e la loro copertura non riguarda la maggior parte della Regio Beta. Il prossimo importante passo avanti nella conoscenza delle strutture superficiali verrà compiuto probabilmente da *Venus Radar Mapper* che la National Aeronautics and Space Administration prevede di lanciare nel 1988 e che coprirà l'intero pianeta con una risoluzione di 0,2 chilometri.

La presenza di anidride solforosa sulla sommità delle nubi venusiane è evidenziata in queste immagini basate su misurazioni eseguite dallo spettrometro per ultravioletto di *Pioneer Venus*. Le immagini rappresentano la radiazione alla lunghezza d'onda di 207 nanometri, fortemente assorbita dall'anidride solforosa e riflessa, invece, dalle nubi. I dati sono stati raccolti in un periodo di cinque giorni nell'agosto 1984 mentre l'atmosfera del pianeta, che ruota molto più velocemente della superficie, eseguiva una rotazione completa al di sotto della sonda. La prima e l'ultima immagine (*rispettivamente in alto a sinistra e in basso a destra*) riproducono la stessa regione dell'atmosfera. Da questo tipo di misurazioni Larry W. Esposito dell'Università del Colorado a Boulder ha stabilito che l'abbondanza media di anidride solforosa al di sopra delle nubi si è ridotta in misura superiore al 90 per cento da quando la sonda è entrata in orbita attorno a Venere, alla fine del 1978. A quell'epoca l'abbondanza di anidride solforosa era più elevata del previsto. Questo suggerisce che prima dell'arrivo del satellite una grande eruzione vulcanica abbia immesso nell'atmosfera superiore anidride solforosa che da allora starebbe formando acido solforico.

Tra le strutture vulcaniche finora scoperte ne esistono di attive? Non è possibile rispondere a questa domanda sulla base delle sole immagini radar. Per dare una risposta senza essere in grado di osservare direttamente le eruzioni è necessario conoscere la composizione della crosta e dell'atmosfera venusiane.

Le prime allettanti informazioni sulla chimica della crosta sono giunte nel 1975 dai moduli di atterraggio *(lander)* delle sonde *Venera 8, 9* e *10* che, dotati di rivelatori di raggi gamma, hanno misurato l'abbondanza nella crosta di potassio, uranio e torio radioattivi. Anche se i luoghi di atterraggio erano molto diversi, sembra che gli elementi radioattivi siano presenti su Venere in quantità paragonabili ai livelli osservati nelle rocce della superficie continentale della Terra. Dal momento che il decadimento radioattivo è la fonte più importante del calore interno della Terra, i dati delle sonde Venera portano a ritenere che la quantità di calore che si sviluppa all'interno di Venere sia approssimativamente analoga a quella prodotta all'interno del nostro pianeta.

Mediante l'altimetro radar a bordo del modulo orbitale di *Pioneer Venus* sono state realizzate mappe della topografia di Venere. Mentre il satellite descrive un'orbita intorno a Venere, l'altimetro misura la distanza dalla superficie; sottraendola dalla distanza dal centro di massa del pianeta, nota da un calcolo preciso dell'orbita, si ottiene il raggio in ogni punto. Sulle mappe i rilievi più elevati sono in giallo e in rosso, i minori in blu intenso. La maggior parte della superficie venusiana è piatta. Mentre il 35 per cento della superficie terrestre è continentale e il 65 per cento è costituito da fondo oceanico, Venere possiede solo due strutture di tipo continentale, che occupano meno del 5 per cento della superficie. Una di queste, Terra Ishtar, compare vicino al polo settentrionale sul globo (*a sinistra*). (La regione polare è vuota perché l'orbita della sonda non passa sopra i poli.) L'area in rosso di Ishtar è Maxwell, il rilievo montuoso più elevato di Venere, che si innalza di oltre 11 000 metri rispetto all'elevazione media. Maxwell è forse di origine vulcanica, ma probabilmente non è attivo. La zona giallo-verde a sud-ovest di Ishtar, al margine sinistro del globo, è la Regio Beta, dove probabilmente è in corso attività vulcanica: nell'immagine prospettica ingrandita (*in alto*) vi è l'esteso altopiano con un gruppo di guglie. Le due vette maggiori di Beta, Mons Theia e Mons Rhea, sono alte più di 4500 metri. A sud di Beta (*a sinistra nell'immagine*) si estende la Regio Phoebe, che è caratterizzata anch'essa da strutture di tipo vulcanico. L'immagine ingrandita in alto è stata realizzata dallo US Geological Survey; l'immagine globale è invece del Jet Propulsion Laboratory.

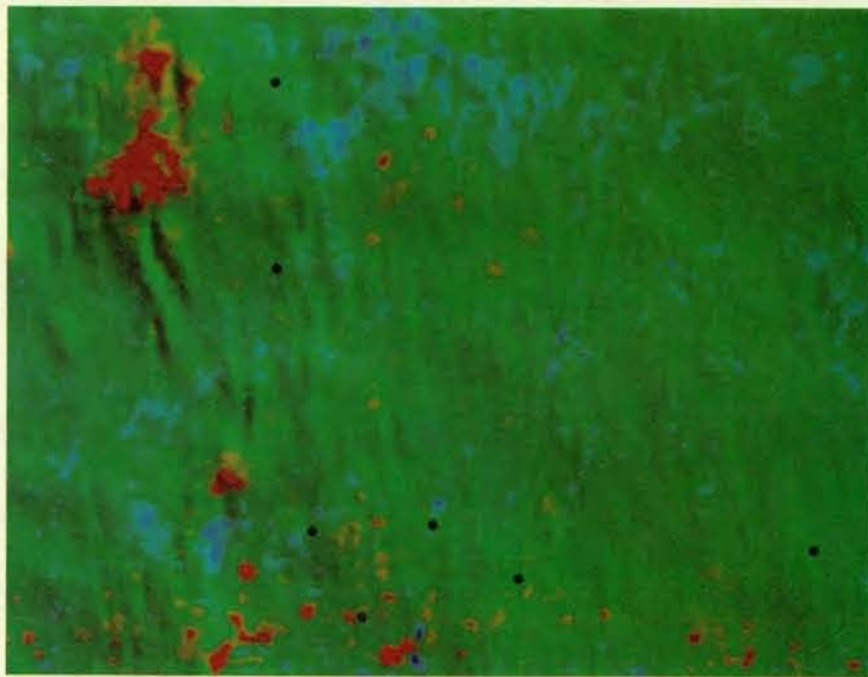
Il calore sviluppato dal decadimento radioattivo, però, deve in qualche modo uscire dall'interno del pianeta. Il sistema più efficiente è costituito da qualche forma di vulcanismo, dove l'espressione, nel senso più generale, indica qualsiasi flusso convettivo di materiale caldo verso la superficie. Sulla Terra molto più della metà della perdita di calore ha luogo lungo le dorsali medio-oceaniche, dove le zolle che costituiscono la litosfera divergono e la lava sgorga dall'astenosfera fluida formando nuova crosta. Una seconda forma di vulcanismo è rappresentata dai vulcani di arco insulare associati a fosse oceaniche, i luoghi dove due zolle collidono e una è subdotta nell'astenosfera. Tuttavia, come hanno osservato Raymond E. Arvidson della Washington University e altri, le mappe radar di Venere non presentano traccia di una rete globale di dorsali e fosse. Può darsi, come ha ipotizzato Don L. Anderson del California Institute of Technology, che la litosfera venusiana sia più sottile di quella terrestre, oltre che più calda e meno densa, e che quindi galleggi troppo per essere subdotta nell'interno fluido del pianeta.

In mancanza di dati certi a favore di una tettonica delle zolle, William M. Kaula e Lynn M. Muradian dell'Università della California a Los Angeles hanno concluso che su Venere la perdita di calore può avvenire attraverso due meccanismi: l'eruzione di vulcani isolati di «punto caldo» come quelli delle Hawaii, che non sono associati a margini di zolla, e il trasferimento di calore per conduzione in zone di crosta sottile e sollevata a «forma di duomo». La struttura a duomo fratturata e a guglie della Regio Beta fa pensare che in questa zona abbiano avuto luogo fenomeni di entrambi i tipi. In generale sembra molto improbabile che la conduzione nelle rocce, un processo molto lento, sia l'unico sistema di raffreddamento. La necessità di un meccanismo di raffreddamento adeguato costituisce una valida considerazione a favore dell'esistenza su Venere di un vulcanismo attivo, ma non è conclusiva.

L'immagine radar della Regio Beta (con una risoluzione orizzontale di circa due chilometri) fa pensare che Mons Theia e Mons Rhea siano vulcani a scudo situati ai fianchi di una frattura (*rift*) rettilinea della crosta. Realizzata alla lunghezza d'onda di 12,6 centimetri con il radiotelescopio da 300 metri dell'Osservatorio di Arecibo, l'immagine mostra le variazioni di scabrosità superficiale: le zone più chiare sono relativamente scabre e non alterate dagli agenti meteorologici, mentre quelle più scure sono relativamente regolari. Theia è la formazione circolare in basso; si attribuisce la sua luminosità alla presenza di colate laviche geologicamente recenti. Si crede che le strutture rettilinee a nord di Rhea siano un *rift* anche perché sono parallele a un canyon scoperto in studi topografici. La morfologia suggerisce che la crosta di Beta sia stata spinta verso l'alto da un sollevamento regionale che ha prodotto il *rift* e i vulcani.

Il vulcanismo ha anche un'altra funzione: lo scarico dei gas dall'interno del pianeta. Lo scarico dei vulcani terrestri contiene anidride carbonica, azoto, vapore acqueo, gas di zolfo e gli isotopi argo 40 ed elio 4 dei rispettivi gas nobili, che sono prodotti dal decadimento radioattivo. Tutti questi gas sono stati rilevati nell'atmosfera venusiana da *Pioneer Venus* e dai moduli di discesa di *Venera 11, 12, 13* e *14*.

In realtà si ritiene che su Venere come sulla Terra, la fuoriuscita di gas dall'interno del pianeta avvenuta nelle prime fasi della sua storia sia stata la più importante fonte di materiale da cui ha avuto origine l'atmosfera. Le due atmosfere, però, sono molto diverse: quella di Venere è per il 96 per cento anidride carbonica mentre quella della Terra è per il 78 per cento azoto. La differenza deriva in parte dalla presenza sul nostro pianeta degli oceani, che hanno estratto dall'atmosfera quasi tutta l'anidride carbonica immagazzinandola in forma di carbonati nella crosta. Gli oceani di Venere, se mai sono esistiti, sarebbero evaporati completamente da molto tempo, e l'idrogeno sfuggito nello spazio.

Confrontando le due atmosfere è necessario quindi inserire nel computo dell'anidride carbonica terrestre i carbonati crostali e anche tener conto del fatto che l'atmosfera di Venere ha una massa 90 volte maggiore. La quantità totale di anidride carbonica e di azoto presente su Venere risulta allora inferiore a quella

La riflettività della superficie venusiana alle onde radio di 17 centimetri varia considerevolmente con la posizione indicando che anche la composizione superficiale varia. La mappa copre la zona da 40 gradi di latitudine nord a 20 gradi sud e da 270 a 340 gradi di longitudine est. La massima riflettività è in rosso, la minima in blu. L'area in rosso, in alto a sinistra, è la zona centrale della Regio Beta; Mons Theia presenta una delle più elevate riflettività finora osservate. Una spiegazione plausibile è che la sua superficie sia costituita da rocce vulcaniche con notevoli inclusioni di pirite, un ottimo conduttore contenente zolfo. Sfortunatamente nessuno dei luoghi di discesa delle sonde Venera (*punti in nero*) si trova in una regione a elevata riflettività. La mappa è stata realizzata da Peter G. Ford e Gordon H. Pettengill del MIT.

terrestre del 30 per cento circa. La concentrazione di argo 40 è circa un terzo di quella che si ha sulla Terra, mentre l'abbondanza nella crosta della sorgente di questo isotopo, il potassio 40, è circa uguale sui due pianeti. Queste indicazioni fanno pensare che essi abbiano subìto fuoriuscite di gas di entità paragonabile e quindi livelli analoghi di vulcanismo nel corso della propria storia. Ciononostante la presenza di anidride carbonica, di azoto e di argo 40 non indica che la perdita di gas abbia avuto luogo recentemente; questi gas sono infatti relativamente stabili e possono sopravvivere nell'atmosfera per periodi geologici. Per avere indicazioni più valide dell'esistenza di attività vulcanica in tempi vicini è necessario considerare le nubi e i gas di zolfo che le producono.

Anche su Venere lo zolfo è un costituente secondario dell'atmosfera (i precursori gassosi delle nubi ne formano lo 0,02 per cento e le particelle stesse delle nubi soltanto lo 0,00002 per cento) eppure le nubi hanno un effetto notevole sul clima del pianeta, perché riflettono quasi l'80 per cento della luce solare incidente, soprattutto alle lunghezze d'onda del giallo e del rosso. Di conseguenza Venere assorbe dal Sole una quantità molto inferiore di energia rispetto alla Terra pur trovandosi più vicino. Dell'energia che non viene riflessa nello spazio i due terzi si depositano nelle nubi, che assorbono nelle lunghezze d'onda dell'ultravioletto e dell'infrarosso vicino; solo un terzo arriva all'atmosfera inferiore e alla superficie. (Senza le nubi quindi la temperatura superficiale sarebbe ancora più elevata.) Sulla Terra avviene invece l'inverso: due terzi dell'energia solare incidente sono assorbiti in superficie.

Le indicazioni del fatto che il componente principale delle nubi sia acido solforico concentrato - 75 per cento in massa - sono indirette, ma convincenti. Lo spettro di riflessione nel visibile e nell'infrarosso delle particelle delle nubi è molto simile a quello dell'acido solforico. Gli studi sulla polarizzazione della luce solare riflessa dalle nubi indicano che le particelle sono di forma sferica, e quindi liquide, e che hanno un indice di rifrazione elevato, pari a 1,44. Questi risultati escludono quasi tutti i possibili candidati diversi dall'acido solforico, e in particolare l'acqua che ha un indice di rifrazione di 1,33 e alle temperature prevalenti nella regione inferiore delle nubi evapora. Inoltre la concentrazione nelle nubi sia di anidride solforosa, sia di vapore acqueo diminuisce rapidamente all'aumentare dell'altezza, e questo induce a ritenere che le due sostanze subiscano reazioni chimiche le quali alla fine producono acido solforico (H_2SO_4).

L'acido solforico è un componente ben noto dell'atmosfera terrestre, presente in forma diluita nella pioggia acida e anche in forma concentrata (come su Venere) in uno strato molto sottile delle nubi stratosferiche. Esso viene prodotto a partire da anidride solforosa, così come da acido solfidrico (H_2S), solfuro di dimetile ($(CH_3)_2S$) e solfuro di carbonile (OCS). La fonte principale di anidride solforosa è l'uso di combustibili fossili, mentre i tre gas ridotti sono prevalentemente sottoprodotti metabolici di vari solfobatteri. Ma l'anidride solforosa e due dei gas sulfurei ridotti (acido solfidrico e solfuro di carbonile) sono anche effluenti comuni dei vulcani terrestri, insieme ad acido cloridrico e acido fluoridrico.

Tutti questi effluenti vulcanici sono stati rilevati nell'atmosfera venusiana. Poiché sono estremamente reattivi e quindi con una vita molto breve, e poiché su Venere non sembrano esservi forme di vita, si potrebbe pensare che la loro stessa presenza costituisca una prova dell'esistenza di attività vulcanica in corso. Verso la fine degli anni sessanta, però, John S. Lewis, ora all'Università dell'Arizona, ha proposto che questi gas derivino dalla «cottura» di rocce superficiali dovuta al calore intenso. I gas si aggiungono all'atmosfera, secondo Lewis, alla stessa velocità alla quale ne vengono asportati dalle reazioni con la superficie, cosicché il flusso netto di un composto in entrata o in uscita dall'atmosfera è nullo. In altre parole i gas di zolfo, l'acido cloridrico e l'acido fluoridrico dell'atmosfera sono in equilibrio chimico con i minerali della crosta.

Secondo il modello di Lewis il rapporto di mescolanza, cioè la concentrazione atmosferica, dell'ossigeno molecolare è determinato da una reazione reversibile all'equilibrio nella quale ossido ferroso (FeO) in forma minerale e solfato di calcio ($CaSO_4$) in forma di anidrite reagiscono con anidride carbonica e producono calcite ($CaCO_3$), pirite (Fe_2S) e ossigeno. Il rapporto di mescolanza previsto per l'ossigeno, che è molto piccolo, determina lo stato di ossidazione dei gas sulfurei. Di conseguenza il gas di zolfo prevalente sarebbe il solfuro di carbonile ridotto a una concentrazione di circa 600 parti per milione. L'acido solfidrico, un altro composto ridotto, avrebbe un rapporto di mescolanza di circa 130 parti per milione; quello dell'anidride solforosa sarebbe di sole 16 parti per milione.

Da quando Lewis ha proposto il suo modello all'equilibrio, nuovi dati, alla cui interpretazione ho contribuito, hanno rivelato che quelle previsioni sono sbagliate, almeno per quanto riguarda i gas di zolfo. Il valore misurato del rapporto di mescolanza dell'anidride solforosa è di 150 parti per milione, quasi 10 volte quello all'equilibrio. Le misurazioni più recenti, compiute nel 1982 dai gascromatografi a bordo delle sonde *Venera 13* e *Venera 14*, collocano l'abbondanza totale di solfuro di carbonile e di acido solfidrico intorno a 150 parti per milione, un valore molto inferiore al loro totale all'equilibrio di 730 parti per milione. Le sonde precedenti, *Venera 11*, *Venera 12* e *Pioneer Venus*, avevano riscontrato livelli di tali composti ancora inferiori, e quindi un'importanza relativa dell'anidride solforosa ancora maggiore.

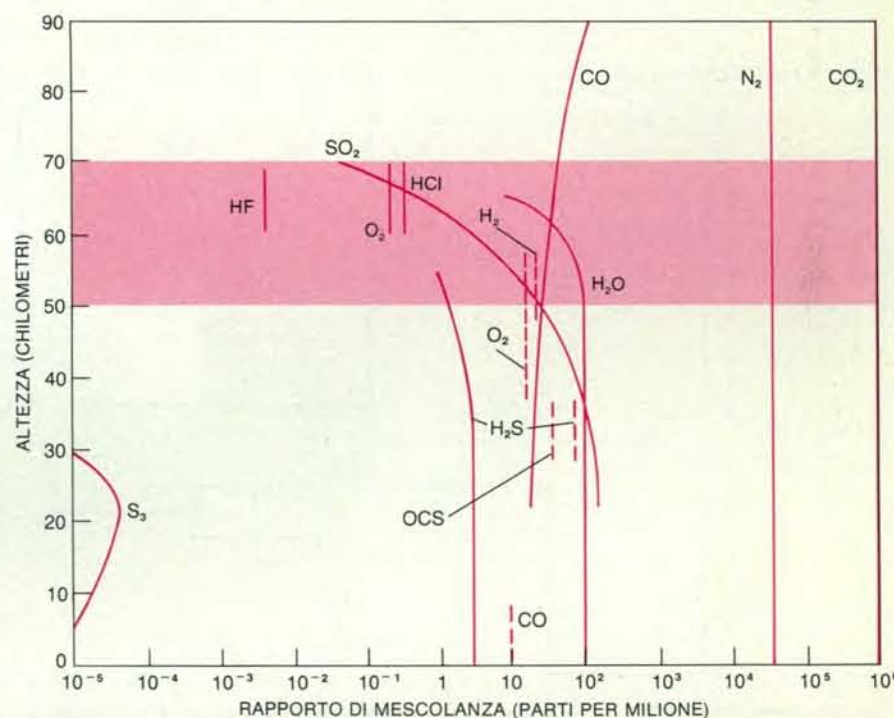
Il modello all'equilibrio non è in grado di spiegare i livelli elevati di anidride solforosa. Oggi è chiaro che le abbondanze dei gas sulfurei nell'atmosfera venusiana non sono determinate da reazioni reversibili con la crosta, che procedono alla medesima velocità nei due sensi. Lo zolfo viene invece immesso nell'atmosfera da una sola reazione, subisce una serie di trasformazioni fino ad acido solforico che, dopo un arco di tempo geologico, viene sottratto all'atmosfera da una reazione di tipo diverso con la crosta.

Su Venere, il ciclo dello zolfo mantiene in equilibrio due regimi chimici opposti, uno fotochimico e uno termochimico. Le reazioni fotochimiche aumentano lo stato di ossidazione dello zolfo e producono le nubi di acido solforico. L'ossigeno proviene dalla decomposizione dell'anidride carbonica in monossido di carbonio e ossigeno da parte della luce ultravioletta all'interno e al di sopra delle nubi. Vicino alla superficie temperatura e densità elevate portano a reazioni termochimiche il cui risultato è una riduzione netta di zolfo. Queste reazioni rigenerano i precursori gassosi delle nubi.

Il ciclo si può suddividere in tre parti: i sottocicli atmosferici «veloce» e «lento» e un sottociclo geologico. Il sottociclo veloce comincia nell'atmosfera media e superiore (sopra la base delle nubi a 50 chilometri di quota) con l'ossidazione di anidride solforosa ad acido solforico per azione della luce ultravioletta. Le reazioni di ossidazione sono catalizzate da composti di cloro e di idrogeno provenienti dalla fotodissociazione di acido cloridrico. L'acido solforico affonda poi nell'atmosfera inferiore calda al di sotto delle nubi ed evapora. L'anidride solforica che ne deriva reagisce per via termochimica con il monossido di carbonio e rigenera anidride carbonica e solforosa. Una molecola di zolfo completa un giro del sottociclo veloce in un anno circa.

Nel sottociclo lento all'interno e al di sopra della regione nuvolosa si formano acido solforico e zolfo elementare per ossidazione fotochimica di acido solfidrico e solfuro di carbonile; nell'atmosfera inferiore questi due composti allo stato gassoso sono ossidati a zolfo e forse ad anidride solforosa per azione delle radiazioni dell'ultravioletto vicino trasmesse dalle nubi. La formazione di zolfo elementare è una caratteristica importante di questo sottociclo, perché spiega l'assorbimento da parte delle nubi della radiazione nella regione ultravioletta dello spettro. Alla fine acido solforico, anidride solforosa e zolfo elementare vengono

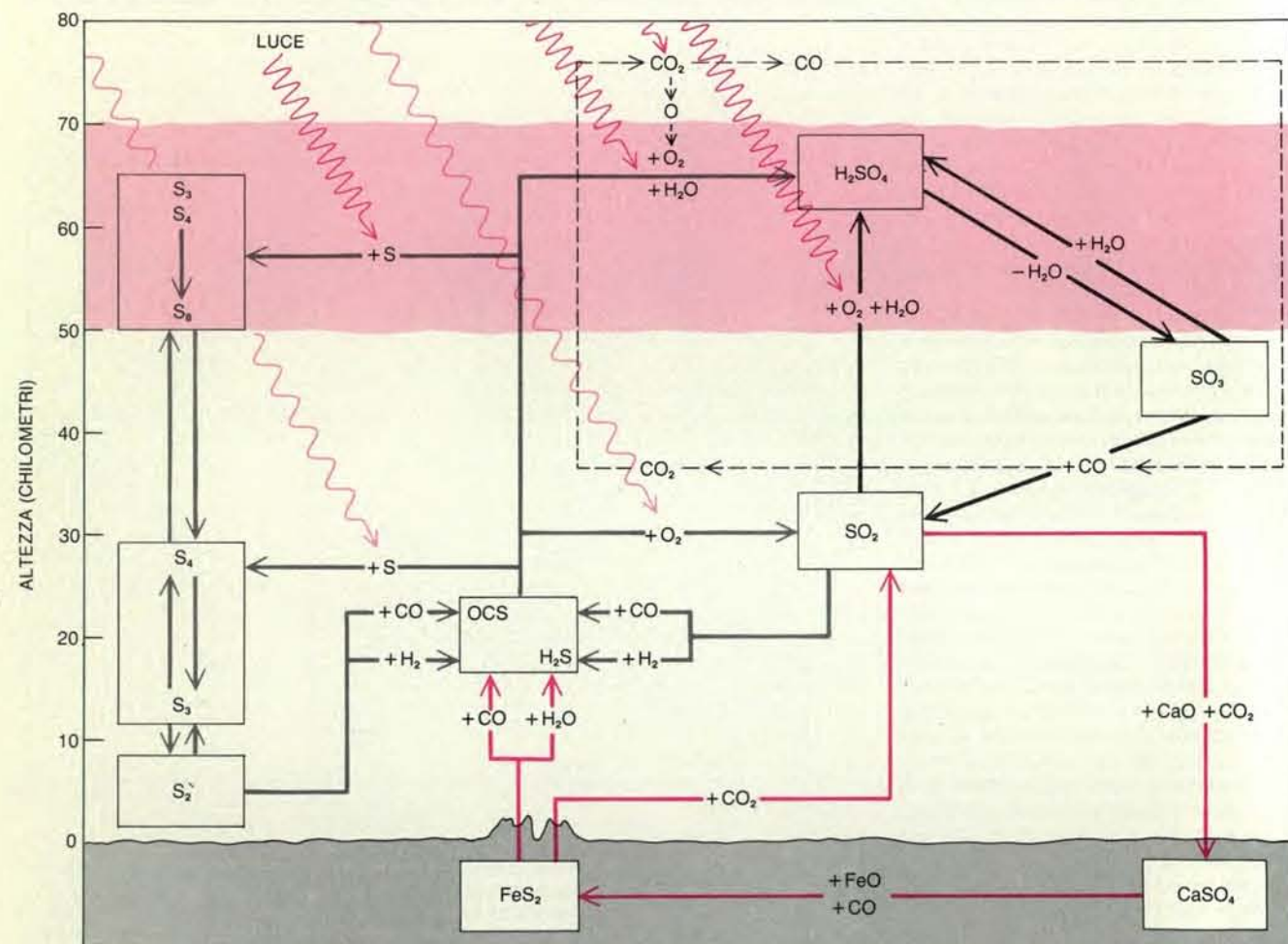
ELEMENTO	VENERA 13	VENERA 14
MAGNESIO (MgO)	11,4 ± 6,2	8,1 ± 3,3
ALLUMINIO (Al_2O_3)	15,8 ± 3,0	17,9 ± 2,6
SILICIO (SiO_2)	45,1 ± 3,0	48,7 ± 3,6
POTASSIO (K_2O)	4,0 ± 0,6	0,2 ± 0,1
CALCIO (CaO)	7,1 ± 1,0	10,3 ± 1,2
TITANIO (TiO_2)	1,6 ± 0,5	1,3 ± 0,4
MANGANESE (MnO)	0,2 ± 0,1	0,2 ± 0,1
FERRO (FeO)	9,3 ± 2,2	8,8 ± 1,8
ZOLFO (SO_3)	1,6 ± 1,0	0,9 ± 0,8
ALTRI	3,9	3,6

La composizione della superficie di Venere è stata determinata in due punti dagli strumenti a fluorescenza X a bordo delle sonde *Venera 13* e *Venera 14*. Nel calcolo dell'importanza relativa dei vari elementi (espressa in percentuale del peso totale) si è presunto che ciascuno fosse presente in forma di ossido. Le concentrazioni sono analoghe a quelle riscontrate in certi basalti vulcanici terrestri, tranne per lo zolfo che è più abbondante su Venere. Su questo pianeta però questo elemento risulta molto meno abbondante del calcio, e ciò indica che la maggior parte del calcio si trova in forma ossidata nei silicati e nei carbonati e non nei solfati. Si ritiene che l'ossido di calcio stia sottraendo anidride solforosa all'atmosfera; i due composti reagiscono spontaneamente allorquando il livello dell'anidride solforosa supera il punto di equilibrio.

La composizione dell'atmosfera venusiana è stata analizzata mediante telescopi installati a terra e con varie sonde spaziali, le ultime delle quali sono *Venera 13* e *Venera 14* (*linee tratteggiate*). Il grafico rappresenta su scala logaritmica i rapporti di mescolanza, ossia le concentrazioni, di alcuni composti in funzione dell'altezza; l'anidride carbonica (CO_2) per esempio, che costituisce il 96 per cento dell'atmosfera, è oltre 10 volte più abbondante dell'azoto (N_2) e circa 10 000 volte più abbondante dell'anidride solforosa (SO_2), il cui rapporto di mescolanza è pari a circa 150 parti per milione. L'interruzione delle linee indica la mancanza di misurazioni oltre una particolare fascia di altezze e non l'assenza dei composti corrispondenti. Le ultime missioni Venera hanno riscontrato livelli di acido solfidrico (H_2S) e di solfuro di carbonile (OCS) superiori a quelli rilevati in precedenza. I valori relativi all'idrogeno (H_2) e all'ossigeno (O_2) sono limiti superiori; la presenza di queste molecole non è stata in realtà rivelata. Anidride solforosa e vapore acqueo presentano diminuzioni d'abbondanza parallele nella regione delle nubi, indicative del fatto che reagiscono e formano particelle di acido solforico. L'atmosfera venusiana contiene anche quantità rilevabili di acido clóridrico (HCl) e fluoridrico (HF).

ridotti dall'idrogeno molecolare e dal monossido di carbonio formando di nuovo acido solfidrico e solfuro di carbonile nell'atmosfera inferiore calda. Il compimento del sottociclo lento richiede probabilmente circa dieci anni, perché lo zolfo resta intrappolato per alcuni anni in quello veloce prima di venire ridotto.

Il rapporto di mescolanza totale dei composti dello zolfo e, quindi, lo spessore delle nubi di acido solforico sono determinati nel sottociclo geologico dall'interazione tra atmosfera e crosta, un processo sul quale si dispone oggi di sufficienti informazioni. Tra i nuovi dati i più importanti provengono dagli apparecchi a fluorescenza X a bordo dei moduli di discesa delle sonde *Venera 13* e *Venera 14*. Con questi dati Yu. A. Surkov e collaboratori dell'Istituto Vernadsky hanno di recente stabilito per la prima volta la composizione in elementi della crosta. Le rocce della superficie di Venere contengono una quantità relativamente elevata di zolfo, ma per il resto le abbondanze dei loro elementi sono simili a quelle osservate in certi basalti terrestri.

Il sottociclo geologico comincia con la produzione di gas di zolfo, soprattutto solfuro di carbonile e acido solfidrico, per reazioni termochimiche della pirite, un minerale contenente zolfo, con anidride carbonica, acqua e monossido di carbonio. Una volta nell'atmosfera i gas attraversano più volte i sottocicli atmosferici veloce e lento. Il risultato dei legami fotochimici tra questi è una conversione netta dei solfuri allo stato gassoso in anidride solforosa, che si accumula quindi nell'atmosfera oltre il livello di equilibrio termochimico. L'anidride solforosa viene rimossa dall'atmosfera e immagazzinata sotto forma di solfato di calcio nella crosta tramite una reazione con anidride carbonica e ossido di calcio. Dopo centinaia o anche milioni di anni il solfato di calcio viene sepolto; poi reagisce con l'ossido di ferro rigenerando la pirite e chiudendo il sottociclo.

Surkov e collaboratori hanno trovato una grande quantità di calcio nelle rocce della superficie venusiana. La concentrazione di zolfo è però molto inferiore, e questo indica che la maggior parte del calcio compare in forma ossidata nei silicati e nei carbonati, e non nei solfati. Dal momento che l'ossido di calcio è disponibile e che il rapporto atmosferico di mescolanza dell'anidride solforosa supera il livello all'equilibrio, tra le due sostanze deve svolgersi una reazione spontanea che porta alla formazione di solfato di calcio. Non vi è dubbio, quindi, che attualmente stia avvenendo una sottrazione di anidride solforosa dall'atmosfera da parte del «pozzo» di ossido di calcio. Si può dunque spiegare la presenza di anidride solforosa a concentrazioni 10 volte superiori a quella di equilibrio solo con un'immissione geologicamente molto recente di gas sulfurei.

Il ciclo dello zolfo responsabile della produzione delle nubi di Venere è costituito da tre sottocicli. Il sottociclo geologico (*linee in colore*) comincia con la reazione termochimica, nel sottosuolo o alla superficie di colate laviche, della pirite vulcanica (FeS_2), che forma anidride solforosa e i composti sulfurei gassosi H_2S e OCS. Nel sottociclo atmosferico lento (*linee in grigio*) e in quello veloce (*linee in nero*) le reazioni fotochimiche generate dalla radiazione ultravioletta all'interno delle nubi e dalla radiazione violetta vicina al di sotto delle nubi ossidano i gas di zolfo formando così le particelle che costituiscono le nubi, ossia acido solforico (H_2SO_4) e varie forme di zolfo elementare (S). L'ossigeno proviene dalla fotodissociazione dell'anidride carbonica. Nell'atmosfera inferiore l'evaporazione e la riduzione termochimica da parte del monossido di carbonio (CO) e dell'idrogeno distruggono il materiale che costituisce le nubi e rigenerano i precursori gassosi di queste. La fotoossidazione porta a una conversione netta dell'acido solfidrico e del solfuro di carbonile in anidride solforosa, che reagisce in superficie con ossido di calcio (CaO) producendo solfato di calcio ($CaSO_4$). Lo zolfo viene così restituito alla crosta.

Nel marzo 1982 il modulo di discesa sovietico *Venera 14* ha fotografato questa pianura venusiana. La temperatura media alla superficie di Venere è di 460 gradi centigradi, l'atmosfera ha una pressione 100 volte superiore a quella della Terra ed è estremamente corrosiva. Ciononostante gli strumenti delle sonde *Venera 13* e *Venera 14* hanno funzionato per diverse ore e hanno trasmesso una serie di fotografie.

La fonte di zolfo ipotizzata, la pirite, è un minerale secondario comune nelle rocce vulcaniche terrestri. Le reazioni che producono gas sulfurei potrebbero aver luogo alla superficie di colate laviche ricche di pirite oppure al di sotto della superficie, nel qual caso questi gas sarebbero espulsi da eruzioni vulcaniche. In ogni caso il meccanismo che garantisce un flusso costante di gas sulfurei nell'atmosfera è di tipo vulcanico.

Negli ultimi tempi Pettengill e Peter G. Ford del MIT hanno ottenuto indicazioni indirette dell'esistenza di estesi affioramenti di pirite nella Regio Beta, dove le mappe radar avevano già rivelato la presenza di strutture vulcaniche. Le osservazioni compiute dal modulo orbitale di *Pioneer Venus* hanno dimostrato che la riflettività della superficie venusiana alle onde radio con una lunghezza d'onda di 17 centimetri varia ampiamente, da un minimo del 3 per cento a un massimo del 40 per cento. Una zona con una riflettività particolarmente elevata è il Mons Theia, nella Regio Beta; Pettengill e Ford sostengono che la composizione di questo rilievo montuoso debba essere decisamente diversa da quella dei bassopiani scarsamente riflettenti che la circondano. Una riflettività elevata implica un'alta conducibilità elettrica, e i tipi di roccia dotati della conducibilità necessaria a spiegare la riflettività di Mons Theia sono decisamente pochi. Una roccia con sostanziali inclusioni di pirite, un efficiente conduttore, rappresenta la possibilità più verosimile. In confronto alla Regio Beta le zone di discesa di *Venera 13* e *14* presentano una riflettività molto bassa; ciò suggerisce che lo zolfo trovato da queste sonde fosse per lo più in forma di solfato di calcio, un cattivo conduttore, e non di pirite.

La pirite superficiale della Regio Beta potrebbe essere il prodotto di colate laviche recenti, ma potrebbe anche essere affiorata per alterazione, dovuta ad agenti meteorologici, di depositi più antichi. Poiché le nubi impediscono l'osservazione diretta di colate laviche attive o di pennacchi di gas e polvere, è impossibile «dimostrare» che su Venere siano in corso eruzioni vulcaniche, come invece hanno dimostrato le spettacolari fotografie di Io riprese dal *Voyager*. Sulla Terra, però, le grandi eruzioni come quella di El Chichón del 1982 lanciano grandi quantità di particelle di caligine e gas sulfurei nell'atmosfera superiore, dove rimangono per mesi o addirittura anni. Se su Venere sono in corso grandi eruzioni, dovrebbe quindi esservene traccia al di sopra delle nubi.

A quanto sembra è proprio così. Larry W. Esposito dell'Università del Colorado a Boulder ha riferito l'anno scorso che lo spettrometro per ultravioletto del modulo orbitale di *Pioneer Venus* ha registrato tra il 1978 e il 1983 una diminuzione del 90 per cento nei livelli di particelle di caligine (acido solforico e anidride solforosa) al di sopra delle nubi, una delle scoperte più sorprendenti di tutta la missione. Nel 1978 i livelli superavano notevolmente quelli previsti dai ricercatori sulla base delle osservazioni compiute da terra nel corso dei quindici anni precedenti; un'analoga sovrabbondanza di particelle di caligine dovrebbe essersi avuta verso la fine degli anni cinquanta. Esposito ha proposto che in entrambi i casi alcune potenti eruzioni vulcaniche abbiano immesso direttamente nell'atmosfera superiore anidride solforosa che poi si sarebbe trasformata in acido solforico il quale sarebbe riaffondato nell'atmosfera inferiore. Gradualmente, quindi, l'abbondanza elevata di anidride solforosa e di particelle di caligine al di sopra delle nubi sarebbe ritornata ai livelli normali.

Non è necessario supporre che i pennacchi vulcanici riescano a spingersi senza diluirsi fino alla sommità delle nubi, a 70 chilometri di altezza, un fenomeno a mio avviso improbabile. Anche se il gas vulcanico caldo di per sé non arrivasse fin sopra le nubi, infatti, l'energia convettiva di una eruzione massiccia si propagherebbe verso l'alto in forma di intense onde gravitazionali di ampiezza progressivamente crescente che finirebbero per frangersi nella regione nuvolosa in modo molto simile alle onde marine su una spiaggia. La concentrazione di anidride solforosa è 500 volte più elevata alla base delle nubi che non alla sommità, e quindi un rimescolamento turbolento causato dal frangersi delle onde gravitazionali potrebbe spiegare l'aumento episodico del livello di anidride solforosa che può verificarsi nell'atmosfera superiore. La spiegazione richiede comunque grandi eruzioni vulcaniche.

Un'altra osservazione compiuta da *Pioneer Venus* indica forse la posizione esatta in cui sono in corso alcune eruzioni. Frederick L. Scarf della TRW Inc. e Christopher T. Russel, dell'Università della California a Los Angeles hanno riferito poco tempo fa che un'antenna sulla sonda ha raccolto alcuni impulsi radio di bassa frequenza che si ritiene siano emessi da fulmini. Gli impulsi sono vistosamente raggruppati sopra diverse strutture superficiali: la Regio Beta, la Regio Atla e la Regio Phoebe, che si trova a sud di Beta. È difficile spiegare questo raggruppamento se si attribuiscono i fulmini a fenomeni di convezione casuale nell'atmosfera. Tutte e tre le regioni, però, sono state riconosciute, per la loro topografia, zone di possibile origine vulcanica, e spesso sulla Terra sono stati osservati fulmini nei pennacchi di vulcani in eruzione. Questa indicazione suggerisce che scariche analoghe abbiano luogo anche su Venere.

Nessuna delle considerazioni fatte è di per sé perfettamente convincente, ma prese tutte insieme costituiscono un'argomentazione molto persuasiva: su Venere esistono vulcani attivi, e questo vulcanismo è un passaggio essenziale del ciclo chimico che produce le nubi. Finora non esiste alcuna indicazione chiara dell'esistenza sul pianeta di una tettonica delle zolle, e quindi i vulcani sono probabilmente punti caldi isolati abbastanza simili al Mauna Loa dell'isola Hawaii. Sembra che i livelli di attività vulcanica dei due pianeti siano grosso modo paragonabili; alcuni ricercatori ritengono anzi che su Venere le eruzioni siano ancora più frequenti. Come minimo si può concludere, dopo decenni di congetture, che il pianeta coperto di nubi, come la Terra, si sta ancora evolvendo, cioè è ancora geologicamente vivo.

Traslocazioni cromosomiche e cancro umano

In una cellula del sistema immunitario i cromosomi possono scambiarsi segmenti di DNA in un processo che attiva geni la cui funzione oncogena è intensificata quando vengono avvicinati ad alcune sequenze genetiche

di Carlo M. Croce e George Klein

Ogni cellula umana contiene oncogèni, geni potenzialmente in grado di provocare il cancro. Gli oncogèni svolgono funzioni normali fino all'insorgere della condizione maligna. Che cosa modifica l'oncogène da componente normale del congegno genetico cellulare a sorgente di trasformazione cancerosa, o neoplastica?.

Nell'ultimo decennio sono stati scoperti meccanismi di differente natura con i quali si può attivare un oncogène (si veda l'articolo *Una base molecolare per il cancro* di Robert A. Weinberg in «Le Scienze» n. 185, gennaio 1984). A volte questa attivazione è dovuta a una «mutazione puntiforme», cioè un piccolo segmento del gene viene alterato da una radiazione o da una sostanza cancerogena. Un altro tipo di attivazione si svolge per «amplificazione», nella quale l'oncogène viene duplicato molte volte cosicché sono presenti, nella stessa cellula, diverse sue copie attive. In questo caso, l'espressione di quel gene ha luogo in modo esageratamente intenso; in altre parole la cellula può produrre in quantità troppo elevata la proteina codificata dal gene e anche una proteina necessaria per il buon funzionamento della cellula può avere effetti cancerogeni quando è prodotta in eccesso (si veda l'articolo *Le proteine degli oncogèni* di Tony Hunter in «Le Scienze» n. 194, ottobre 1984).

Un oncogène potrebbe venir attivato anche mediante incorporazione in un retrovirus (cioè in un virus il cui materiale genetico consta di RNA anziché di DNA). Quando un retrovirus infetta una cellula animale, può catturare da questa cellula un oncogène non attivato che entra a far parte del corredo genetico di quel retrovirus e della sua progenie. Talvolta il processo attiva l'oncogène e così la successiva infezione da parte di quel ceppo di retrovirus può indurre una trasformazione neoplastica in una cellula diversa. Fino a questo momento non è chiaro quale ruolo svolgano nello sviluppo dei tumori umani questi meccanismi di attivazione, dato che solo pochi sono i tumori umani che hanno un oncogène attivato in uno qualsiasi dei suddetti modi.

Il nostro lavoro e quello di altri ricercatori hanno dimostrato che vi è un altro meccanismo, in grado di attivare un oncogène: esso opera in alcuni tumori maligni di cellule del sistema immunitario, chiamate cellule *B*. La funzione primaria di una cellula *B* consiste nel produrre anticorpi (immunoglobuline), molecole che riconoscono altre molecole estranee, gli antigeni, e si legano a esse. L'espressione dei geni che codificano per la produzione degli anticorpi deve avvenire ad alto livello perché la cellula *B* possa svolgere la propria funzione in maniera adeguata. Le sequenze genetiche all'interno dei geni che codificano per la produzione degli anticorpi aumentano l'attività di quei geni nelle cellule *B*. Se un riassestamento dei cromosomi di una di queste cellule (i cromosomi sono filamenti bastoncellari di DNA, contenenti i geni) in un modo o nell'altro giustapponesse una sequenza di questo tipo a un oncogène, allora l'espressione di quell'oncogène verrebbe incrementata. La trasformazione maligna diventerebbe apparentemente una componente primaria della funzione di quella cellula.

Noi abbiamo trovato che effettivamente riassestamenti del genere avvengono nel linfoma di Burkitt, un tumore maligno del sistema immunitario a sviluppo estremamente rapido. Essi conseguono a traslocazioni reciproche tra due cromosomi della cellula *B*: un segmento di ogni cromosoma si stacca e si sposta all'estremità dell'altro cromosoma (*si veda l'illustrazione a pagina 54*). Nella maggior parte di queste traslocazioni un oncogène si porta in una posizione vicina a una delle sequenze che intensificano la produzione di anticorpi; meno spesso l'oncogène rimane al suo posto e la sequenza ad azione stimolante si sposta.

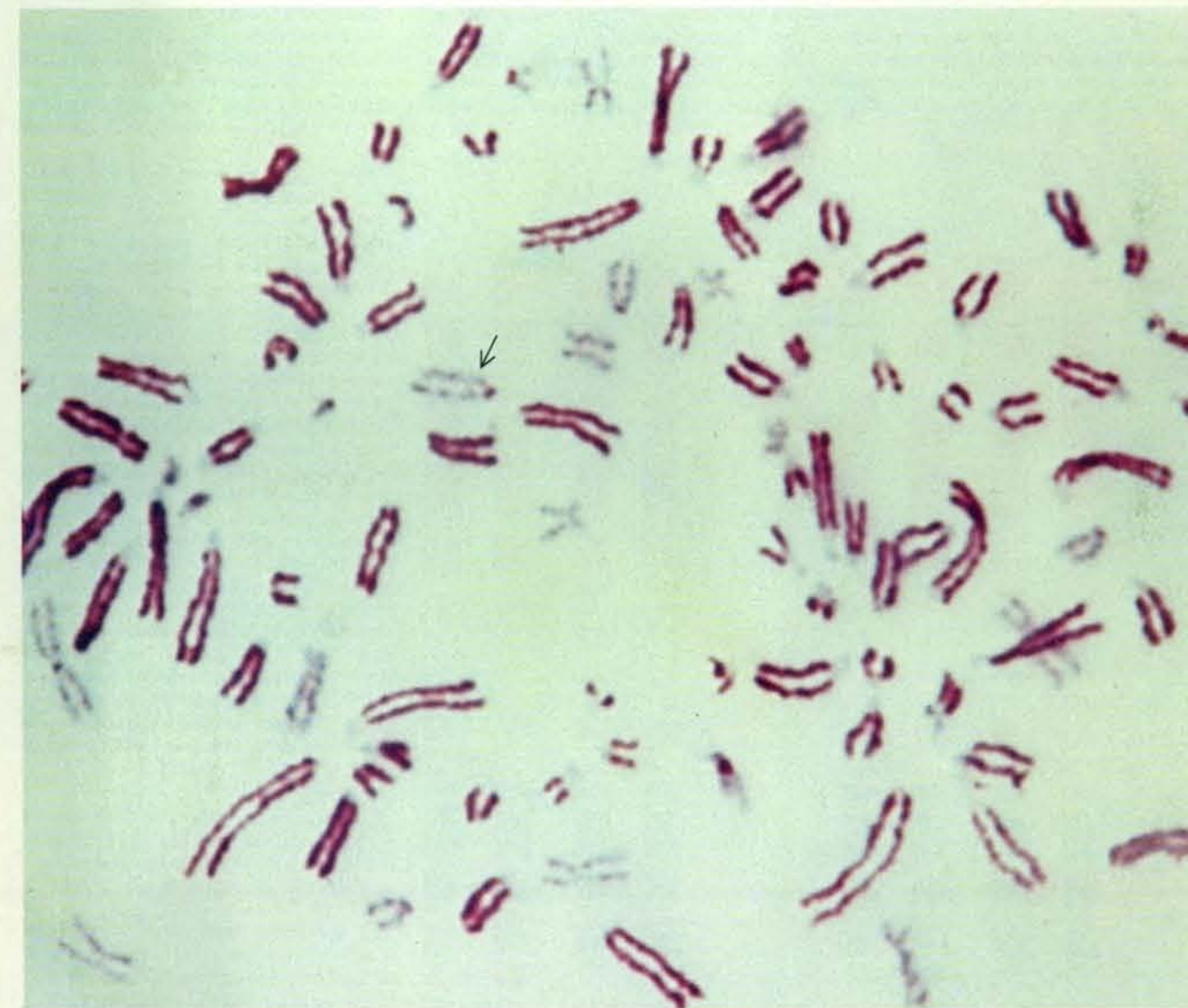
Abbiamo incontrato questo meccanismo nel corso di una ricerca avviata alla fine degli anni settanta per identificare i cromosomi che contengono i geni responsabili della produzione degli anticorpi. Dopo aver localizzato la posizione di questi geni sui cromosomi, abbiamo notato che si trovano sugli stessi cromosomi di cui era già nota la traslocazione nelle cellule del linfoma di Burkitt. La nostra successiva ricerca ha dimostrato che i due segmenti costituiti di materiale genetico che, nel corso della traslocazione, si spostano da un cromosoma all'altro contengono rispettivamente un oncogène e un gene che codifica per una parte della molecola dell'anticorpo.

Per scoprire quali cromosomi contengano l'informazione genetica che codifica per la produzione di anticorpi, uno di noi (Croce) e collaboratori hanno fatto ricorso a una tecnica sperimentale che contempla l'uso di ibridi tra cellule somatiche umane e di topo (le cellule somatiche si contrappongono alle cellule sessuali - cellule uovo e spermatozoi - e sono le cellule corporee).

Le cellule ibride si ottengono mescolando cellule *B* di topo e umane in un mezzo contenente un fattore di fusione chimico o virale, che congiunge cellule dei due tipi. Esse presentano così sia cromosomi umani sia cromosomi di topo e ciascuna può perdere alcuni cromosomi umani durante la divisione cellulare (pur conservando l'intero corredo cromosomico di topo); pertanto, a mano a mano che le cellule si moltiplicano, le successive generazioni possiedono sempre meno cromosomi umani e così, dopo parecchie generazioni, ogni cellula ibrida ne avrà soltanto pochi.

Per sapere quali cromosomi contengono i geni che codificano per parti della molecola di un anticorpo, abbiamo esaminato diversi gruppi di queste cellule. Qualunque cellula che producesse una parte della molecola doveva contenere uno dei cromosomi necessari. Siamo stati in grado di determinare quali cromosomi contengono i geni per una qualsiasi parte particolare della molecola di anticorpo notando quali cromosomi umani erano sempre presenti nelle cellule che producevano quella parte, ma assenti in qualunque cellula che non producesse quella parte.

La molecola di anticorpo consta di quattro catene polipeptidiche, che si legano in due coppie identiche a forma di Y (*si veda l'illustrazione a pagina 55*). La catena più lunga, in ogni coppia, è chiamata catena pesante, quella più corta catena leggera. Ciascuna consiste di due regioni caratteristiche: la regione «variabile» e la regione «costante». La prima riconosce gli antigeni e si lega a essi; la seconda specifica il compito che l'anticorpo deve svolgere (cioè la sua funzione effettrice) dopo aver incontrato l'antigene ed essersi legato a esso. Vi sono numerosi tipi diversi di regione variabile, in quanto gli anticorpi sono estremamente selettivi, cioè ciascuno si lega soltanto a un antigene specifico. Per contro sono presenti soltanto due tipi di regione costante nelle catene leggere (rispettivamente kappa e lambda) e dieci tipi nelle catene pesanti. Capita così che gli anticorpi contro differenti antigeni eseguano lo stesso compito. Ogni cellula *B* matura può produrre soltanto un tipo di anticorpo e i suoi cromosomi contengono DNA che codifica per le regioni variabile e costante, specifiche per quell'anticorpo (si veda l'articolo *La genetica della diversità tra anticorpi* di Philip Leder in «Le Scienze» n. 167, luglio 1982).

Nel 1979, uno di noi (Croce) e collaboratori hanno scoperto che le cellule ibride contenenti il cromosoma umano 14 erano le sole in grado di produrre catene pesanti. È evidente che i geni che codificano per la produzione di queste catene sono situati nel cromosoma 14. Sfruttando analoghe tecniche sperimen-

I cromosomi di cellule ibride, ottenute da cellule umane e di topo, includono sia cromosomi umani (*in colore chiaro*), sia cromosomi di topo (*in colore intenso*). Uno dei cromosomi umani (*freccia*) ha subito una traslocazione, cioè un segmento a una estremità si è staccato ed è stato sostituito da un segmento proveniente da un altro cromosoma. Le traslocazioni possono attivare gli oncogèni, geni che provocano il cancro, trasferendoli in vicinanza di sequenze genetiche «intensificatrici», il cui ruolo consiste nell'aumentare l'attività di certi altri geni presenti sullo stesso cromosoma. Dato che le cellule ibride contengono una parte del corredo genetico umano, ma non tutto, possono venir utilizzate per determinare i cromosomi che codificano per un prodotto umano: qualunque prodotto sintetizzato da una cellula contenente appena un cromosoma umano deve derivare da quel cromosoma. Gli autori dell'articolo hanno fatto ricorso alle cellule ibride per riuscire a identificare i cromosomi contenenti certi oncogèni e per studiare gli effetti di varie traslocazioni sulla regolazione di questi oncogèni.

tali, Jan Erikson, Joanne Martinis e uno di noi (Croce) hanno trovato nel 1981 che il cromosoma 22 codifica per le catene leggere che contengono la regione costante lambda. Nel 1982 O. Wesley McBride e collaboratori al National Cancer Institute e Terence H. Rabbitts e collaboratori del Laboratory of Molecular Biology del Medical Research Council a Cambridge hanno trovato che il cromosoma 2 codifica per le catene leggere che contengono la regione costante kappa.

Questi risultati concordavano bene con quelli di lavori che erano stati effettuati un decennio prima sulle traslocazioni cromosomiche. Nel 1972 George Manolov e Yanka Manolova, all'Università di Lund in Svezia, hanno trovato una irregolarità nei cromosomi di molte cellule che erano state colpite dal linfoma maligno di Burkitt: uno dei cromosomi del quattordicesimo paio (una cellula somatica umana ha 23 coppie distinte di cromosomi) era allungato. I Manolov, notando che una parte del cromosoma, il braccio *q*, aveva una lunghezza anomala, hanno chiamato questo cromosoma $14q^+$.

Successivamente, Lore Zech in collaborazione con uno di noi (Klein) al Karolinska Institut ha trovato che il cromosoma $14q^+$ è il risultato di una traslocazione reciproca: cioè si forma quando un segmento terminale di un cromosoma dell'ottavo paio si stacca e si unisce al cromosoma 14. Un segmento di questo cromosoma subisce una transizione in senso opposto e va a congiungersi con l'estremità del cromosoma 8. Questo cromosoma, così ristrutturato, prende il nome di $8q^-$, perché ha un braccio *q* accorciato. Più di recente altri ricercatori hanno scoperto che, nelle cellule del linfoma di Burkitt, possono aver luogo due altre traslocazioni cromosomiche. Ambedue interessano il cromosoma 8; in un tipo di traslocazione (che ha luogo in circa il 16 per cento dei casi di linfoma di Burkitt), lo scambio reciproco avviene tra i cromosomi 8 e 22. In circa il 9 per cento dei casi esso interessa i cromosomi 2 e 8. Tre dei cromosomi colpiti da queste traslocazioni - il 2, il 14 e il 22 - sono interessati nella produzione di anticorpi.

L'associazione tra linfoma di Burkitt e produzione di anticorpi si è dimostrata presto più stretta. Uno di noi (Croce) assieme a Erikson presso il Wistar Institute of Anatomy and Biology e Janet Finan e Peter C. Nowell della School of Medicine dell'Università della Pennsylvania hanno rilevato che il punto in cui il cromosoma 14 si rompe durante la traslocazione con il cromosoma 8 è situato esattamente all'interno di quella parte di cromosoma 14 che codifica per la catena pesante dell'immunoglobulina. Per questi esperimenti abbiamo fatto ricorso ancora una volta a ibridi di cellule umane e di cellule di topo; in questo caso, in particolare, abbiamo utilizzato cellule del sistema immunitario di topo che erano state derivate da un tipo di cancro noto come plasmacitoma. Ogni cellula ibrida conteneva, oltre al corredo genetico del topo, almeno un cromosoma di cellula di linfoma di Burkitt umano.

Come prevedevamo, le cellule ibride con il cromosoma normale 14 (cioè il cromosoma che non era stato toccato dalla traslocazione) possedevano geni per la produzione di anticorpi; quelli con il cromosoma normale 8 non ne possedevano (*si veda l'illustrazione a pagina 58*). D'altra parte, gli ibridi con un cromosoma 14 interessato in una traslocazione (il cromosoma $14q^+$) contenevano i geni per le regioni costanti delle catene pesanti, ma non quelli per le regioni variabili. Il cromosoma 8 che aveva preso parte alla traslocazione conteneva i geni per le regioni variabili. Questi risultati sono la prova che il cromosoma 14 si rompe tra i geni che codificano per la regione variabile e quelli che codificano per la regione costante della catena pesante e che i geni che codificano per la regione variabile si spostano sul cromosoma 8. Il locus per la catena pesante (cioè quella parte del cromosoma 14 che codifica per la catena pesante) è così direttamente interessato in una delle traslocazioni caratteristiche del linfoma di Burkitt.

A questo punto era chiaro che il meccanismo del linfoma di Burkitt era in qualche modo in relazione con i geni che codificano per la produzione di anticorpi. Ulteriori chiarimenti sulla natura di tale rapporto sono emersi dagli studi sugli oncogèni. Dato che il linfoma di Burkitt colpisce le cellule *B*, il nostro interesse era rivolto in particolare verso l'oncogène *c-myc*, un oncogène umano che è in stretta relazione con l'oncogène *v-myc*, il quale provoca nei polli infettati con il virus della mielocitomatosi aviaria un linfoma delle cellule *B*.

In collaborazione con Riccardo Dalla-Favera e Robert C. Gallo del National Cancer Institute, abbiamo sfruttato la stretta relazione tra geni *myc* umani e aviari per costruire una sonda che avrebbe identificato cellule ibride contenenti l'oncogène umano *c-myc*. Questa sonda era un segmento di DNA umano marcato con isotopi radioattivi, la cui sequenza genica risultava molto simile a quella dell'oncogène *v-myc*.

Per sapere se una cellula ibrida conteneva l'oncogène umano *c-myc*, abbiamo utilizzato un enzima per ridurre in piccoli segmenti il materiale genetico; quindi abbiamo separato questi segmenti in base alle dimensioni ricorrendo al metodo della elettroforesi su gel. In una fase successiva abbiamo esposto il DNA a una soluzione contenente DNA radioattivo che fungeva da sonda. Sia il DNA cellulare sia la sonda erano stati in precedenza denaturati, cioè in ciascuno dei due casi i due filamenti complementari di DNA che costituiscono questa molecola conformata a doppia elica erano stati separati. Dato che la sonda *c-myc* e l'oncogène umano *c-myc* sono pressoché identici, i filamenti marcati della prima si sono ibridati con il *c-myc* cellulare, cioè singoli filamenti del DNA che funge da sonda si sono uniti a filamenti complementari di DNA cellulare. Quando abbiamo eliminato la soluzione contenente il DNA sonda, tutto il DNA sonda che si era ibridato è rimasto dove era. Dopo questo processo, qualunque cellula il cui materiale genetico si era ibridato con la sonda radioattiva ha potuto essere identificata grazie a una banda radioattiva specifica comparsa sulla carta da filtro che tratteneva il DNA legato.

Abbiamo utilizzato la sonda per esaminare un gruppo di cellule ibride di topo e umane allo scopo di determinare la localizzazione cromosomica del gene umano *c-myc*. Abbiamo esaminato in primo luogo le cellule ibride con cromosomi umani normali e abbiamo trovato che il cromosoma umano 8 era presente in tutte le cellule contenenti l'oncogène *c-myc* umano e assente in quelle che ne erano prive; abbiamo concluso che l'oncogène *c-myc* è situato sul cromosoma 8.

Successivamente abbiamo esaminato cellule ibride che contenevano i cromosomi 8 e 14 traslocati, derivati da unioni tra cellule di topo e cellule di linfoma di Burkitt umano. Abbiamo osservato che l'oncogène *c-myc* risiede nel piccolo segmento di cromosoma 8 che trasloca costantemente sul cromosoma 14 nelle cellule del linfoma di Burkitt contenenti la traslocazione tra cromosomi 8 e 14. Questo risultato ha messo in luce che le traslocazioni che interessano l'oncogène *c-myc* svolgono un ruolo fondamentale nello sviluppo del linfoma di Burkitt.

È interessante notare come analoghe traslocazioni cromosomiche specifiche siano state osservate anche nei plasmacitomi di topo da Shinsuke Ohno, Francis Wiener e Jack Spira, che lavoravano al Karolinska Institut con uno di noi (Klein), insieme con Michael Potter e collaboratori del National Cancer Institute.

La traslocazione reciproca tra cromosoma 8 (*in rosso*) e cromosoma 14 (*in verde*) provoca la maggior parte dei casi di linfoma di Burkitt, un cancro delle cellule *B* del sistema immunitario umano. Un segmento di una estremità del cromosoma 8 si stacca (*a*) e si sposta sul cromosoma 14 (*b*). La traslocazione inversa muove un segmento dal cromosoma 14 al cromosoma 8. Con queste traslocazioni reciproche un oncogène del cromosoma 8 finisce per trovarsi vicino a un gene del cromosoma 14 che codifica per la produzione di parte di un anticorpo. Un meccanismo che intensifica la produzione di anticorpi nelle cellule *B* normali attiva quindi quell'oncogène.

La molecola di anticorpo consiste di due coppie identiche di catene proteiche unite in modo da dar luogo a una forma scissa a *Y*. In ogni coppia sono presenti una catena pesante e una catena leggera e ognuna di queste catene presenta una regione variabile e una regione costante. La maggior parte dei casi di linfoma di Burkitt sono provocati da una traslocazione di un oncogène nel locus genetico che codifica per la catena pesante. Altri casi sono provocati da traslocazioni cromosomiche che interessano geni per le regioni costanti della catena leggera.

Il loro studio ha permesso di scoprire che le cellule maligne produttrici di anticorpi, nel topo, portavano una traslocazione caratteristica tra il cromosoma 15 e i cromosomi di topo che hanno o i geni per la catena pesante o i geni per la regione kappa della catena leggera (rispettivamente cromosomi 12 e 6 di topo). Questi risultati hanno suggerito che i geni per le immunoglobuline hanno un ruolo nei plasmacitomi di topo.

Ulteriori esperimenti, realizzati da uno di noi (Croce), da Dalla-Favera e da Gallo, e da Philip Leder della Harvard Medical School, in collaborazione con Stuart A. Aaronson del National Cancer Institute, hanno dimostrato che l'oncogène *c-myc* traslocato sul cromosoma 14 nel linfoma di Burkitt umano può presentarsi in diversi modi.

Esso consiste di tre esoni (segmenti di DNA che vengono trascritti in RNA messaggero(m-RNA), durante il processo che porta all'espressione del gene come proteina), intercalati da due introni (segmenti di DNA che non vengono trascritti in m-RNA e che, pertanto, non si esprimono come proteine). La struttura dell'oncogène è stata analizzata da Dalla-Favera, Gallo e collaboratori e da Rosemary Watt, Giovanni Rovera e uno di noi (Croce) al Wistar Institute. In alcune traslocazioni del linfoma di Burkitt il punto di rottura sul cromosoma 8 è «a monte» dell'intero oncogène *c-myc* e tutti e tre gli esoni di questo oncogène sono traslocati sul cromosoma 14; in altri casi, invece, il punto di rottura è «a valle» del primo esone e solo il secondo e il terzo vengono traslocati (*si veda l'illustrazione a pagina 60*). In questo caso, l'oncogène si unisce «testa contro testa» a uno dei geni della catena pesante, presenti sul cromosoma 14; in altre parole il segmento di DNA traslocato dal cromosoma 8 decorre in una direzione opposta a quella del DNA proveniente dal cromosoma 14.

Altri esperimenti hanno dimostrato che un analogo riassestamento ha luogo in traslocazioni che sono alla base dei plasmacitomi di topo. Essi sono stati realizzati per la prima volta da Michael D. Cole e collaboratori al Medical Center della St. Louis University e, in seguito, da Leder, da uno di noi (Croce) in collaborazione con Kenneth B. Marcu della State University of New York a Stony Brook, e da Jerry Adams e Suzanne Cory del Walter and Eliza Hall Institute of Medical Research a Melbourne. Nei plasmacitomi di topo l'oncogène *c-myc* subisce un riassestamento e si presenta testa contro testa con un gene di catena pesante di immunoglobulina. Non è ancora chiaro, tuttavia, se l'oncogène viene traslocato nel locus per la catena pesante o se, invece, rimane sul cromosoma 15 di topo mentre il locus per la catena pesante viene traslocato vicino a esso.

Malgrado i vari possibili riassestamenti cromosomici che si realizzano nelle cellule del linfoma di Burkitt umano, abbiamo trovato che la proteina prodotta dall'oncogène *c-myc* era, qualitativamente, la stessa. In particolare abbiamo trovato che il primo esone di *c-myc* non codifica per una proteina; la sintesi proteica comincia con il secondo esone. Per-

I cromosomi traslocati di una cellula di linfoma di Burkitt differiscono come lunghezza dai loro corrispondenti cromosomi non traslocati. In questa cellula, uno dei cromosomi dell'ottavo paio ha subito una traslocazione con un cromosoma del quattordicesimo paio. Il cromosoma 8 traslocato si è accorciato, mentre il cromosoma 14, esso pure traslocato, si è allungato.

tanto non erano stati i riassestamenti dell'oncogène *c-myc* durante la traslocazione ad averne attivato le qualità oncogene; l'effetto canceroso della traslocazione non è dovuto a qualche alterazione all'interno del gene.

Se il prodotto di *c-myc* è lo stesso nelle cellule normali e nelle cellule del linfoma di Burkitt, qual è l'effetto oncogeno della traslocazione cromosomica nel linfoma di Burkitt? Forse la traslocazione in qualche modo fa sì che il prodotto dell'oncogène *c-myc*, che in piccole quantità può essere necessario per il funzionamento della cellula, si possa esprimere a livelli anormalmente elevati.

In altre parole, può darsi che la traslocazione renda l'oncogène *c-myc* in grado di sfuggire ai meccanismi che normalmente ne controllano l'espressione. Se questo è il caso, dovrebbe esservi una differenza tra i livelli di espressione del gene traslocato e dell'oncogène normale *c-myc* nella stessa cellula. Cellule con il cromosoma $14q^+$ dovrebbero avere livelli elevati di m-RNA (materiale genetico che rappresenta uno stadio intermedio tra la presenza di un gene su un cromosoma e la sua espressione come proteina), trascritto dal DNA di *c-myc*. Cellule con il cromosoma normale 8 dovrebbero averne, invece, bassi livelli.

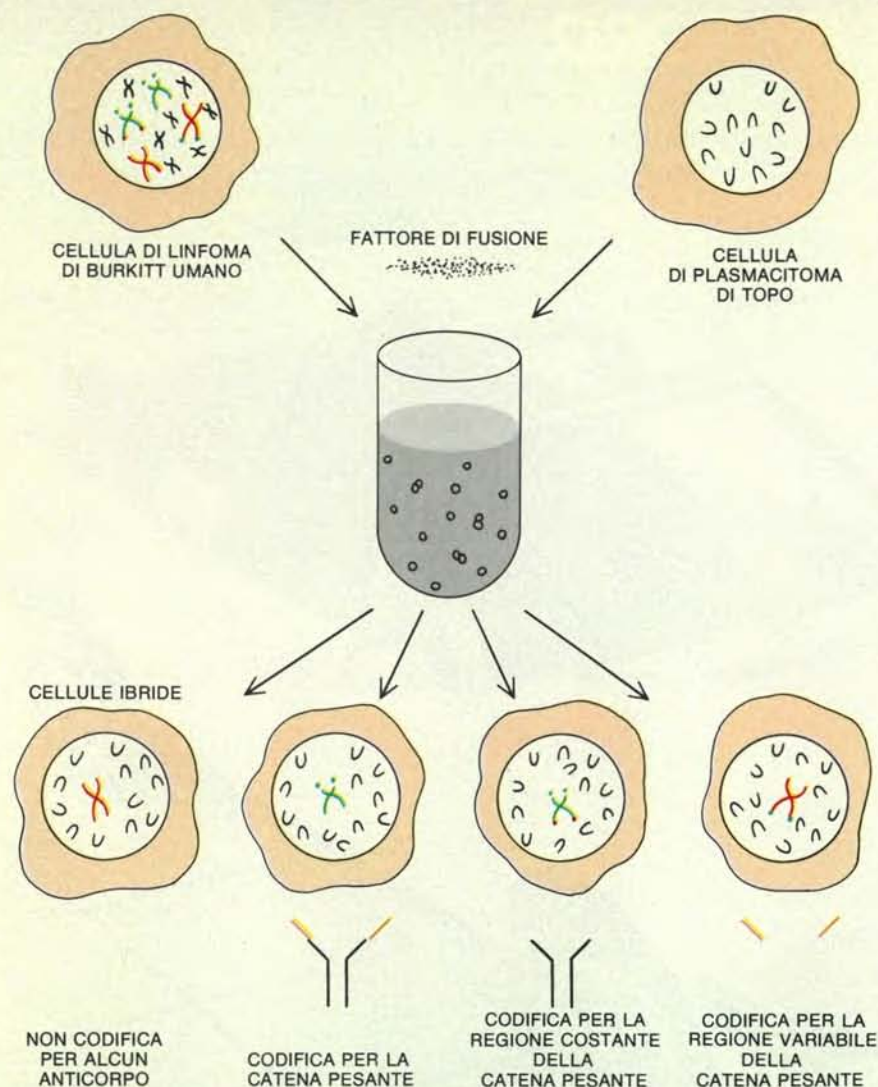
Tenendo presente questa possibilità Kazuko Nishikura e i nostri compagni di lavoro al Wistar Institute hanno avviato altri esperimenti con ibridi di cellule umane e di plasmacitomi di topo. Grazie a un metodo che ci ha permesso di distinguere gli m-RNA trascritti dall'oncogène *c-myc* umano dall'm-RNA trascritto dal *c-myc* di topo, abbiamo trovato che il gene *c-myc* sul cromosoma $14q^+$ si esprime a livelli elevati; invece,il gene *c-myc* sul cromosoma 8 normale è relativamente silente nello stesso tipo di cellula di plasmacitoma.

In esperimenti paralleli abbiamo introdotto un gene *c-myc* sul cromosoma 8 normale in cellule di plasmacitoma di topo. Quel gene era derivato da cellule *B* umane non cancerose e abbiamo trovato che esso, che era stato espresso (anche se a bassi livelli) nelle cellule *B* umane, era completamente silente nelle cellule di plasmacitoma di topo. In altri studi di Adams e Cory hanno trovato che in cellule di plasmacitoma di topo l'oncogène *c-myc* di topo non traslocato non si esprime. Così, mentre un gene *c-myc* normale (non traslocato) viene represso in una cellula di plasmacitoma di topo, un oncogène *c-myc* che viene traslocato sul cromosoma 14 nel locus per la catena pesante sfugge in qualche modo ai meccanismi che normalmente controllano la trascrizione.

Abbiamo anche preso in esame gli m-RNA trascritti nel *c-myc* di cellule di linfoma di Burkitt che presentano una traslocazione variante. In queste cellule, infatti, il primo esone del gene *c-myc* rimane sul cromosoma 8 e gli altri due esoni sono ristrutturati in modo da trovarsi testa contro testa con i geni del cromosoma 14. (Ciascuna di queste cellule aveva, oltre al cromosoma traslocato, un cromosoma 8 normale.) In questi casi la differenza tra m-RNA del gene traslocato e quello del gene presente sul cromosoma 8 normale è relativamente facile da scorgere: il gene traslocato ha subito un riassestamento parziale e, pertanto, gli m-RNA da esso trascritti saranno diversi. In queste cellule Abbas ar-Rushdi e altri nostri collaboratori hanno potuto osservare elevati livelli di m-RNA trascritti dal gene *c-myc* traslocato, ma non dal gene *c-myc* normale.

Questi risultati indicano che l'oncogène *c-myc* va incontro a una «deregolazione» come conseguenza della vicinanza a geni che codificano per anticorpi. Questa conclusione è rafforzata da osservazioni riguardanti due traslocazioni che hanno luogo nel linfoma di Burkitt e che non interessano il cromosoma 14. Una di queste traslocazioni avviene tra il cromosoma 8 e il cromosoma 22, il cromosoma che contiene i geni che codificano per le catene leggere del tipo lambda. L'altra avviene tra il cromosoma 8 e il cromosoma 2, cioè quel cromosoma che contiene i geni che codificano per le catene leggere kappa. In ambedue, come uno di noi (Croce) ha dimostrato in collaborazione con Nowell e con Gilbert Lenoir del Centre International de Recherche sur le Cancer di Lione, il gene *c-myc* rimane sul cromosoma 8, dove viene raggiunto da una sequenza che codifica per la produzione di anticorpi (in un caso dal locus per la catena leggera lambda e nell'altro dal locus kappa). O l'una o l'altra traslocazione attiva l'oncogène rendendolo incapace di reagire ai meccanismi che ne controllano l'espressione. È evidente che l'oncogène *c-myc* non deve muoversi per esprimersi a livelli elevati.

Che cosa è responsabile della deregolazione dell'oncogène *c-myc* che ha luogo in queste traslocazioni? Un'osservazione sperimentale suggerisce una risposta: l'oncogène *c-myc* traslocato del linfoma di Burkitt viene represso nelle cellule ibride ottenute da fibroblasti (cellule del tessuto connettivo) di topo, mentre si esprime a livelli elevati in quelle ottenute da cellule di plasmacitoma (cellule maligne produttrici di anticorpi). Sembra che la traslocazione abbia un effetto oncogeno solo in una cellula che produce anticorpi, cioè una cellula in cui le regioni cromosomiche necessarie per la produzione di anticorpi siano particolarmente attive.

Queste regioni cromosomiche contengono un tipo di sequenza genica che è chiamato sequenza «intensificatrice» (*enhancer*). Sequenze geniche con funzione intensificatrice sembrano accrescere i livelli di trascrizione di certi altri geni sullo stesso cromosoma; esse sono una recente scoperta e poco si sa della modalità del loro funzionamento. Ricercatori nel laboratorio di Kathryn L. Calame dell'Università della California a Los Angeles, in quello di Walter Shaffner dell'Università di Zurigo e di Susu-

Il punto di rottura sul cromosoma 14 che subisce la traslocazione si trova all'interno della regione che codifica per la catena pesante dell'anticorpo, come è dimostrato in questo esperimento. Una cellula di linfoma di Burkitt umano, contenente i cromosomi 8 e 14 normali e traslocati, è stata fusa con una cellula di plasmacitoma di topo (una cellula cancerosa del sistema immunitario di topo). Ogni cellula ibrida ha conservato uno dei cromosomi della cellula umana. Le cellule ibride contenenti il cromosoma 8 normale (*in rosso*) non hanno prodotto anticorpo; quelle con il cromosoma 14 normale (*in verde*) hanno prodotto la catena pesante dell'anticorpo. Il cromosoma 14 interessato nella traslocazione conteneva geni soltanto per le regioni costanti delle catene pesanti e il cromosoma 8, pure interessanto nella traslocazione, conteneva geni per la regione variabile delle stesse catene. È chiaro che, nella traslocazione, il cromosoma 14 si rompe direttamente tra i loci che codificano per le regioni costanti e per le regioni variabili.

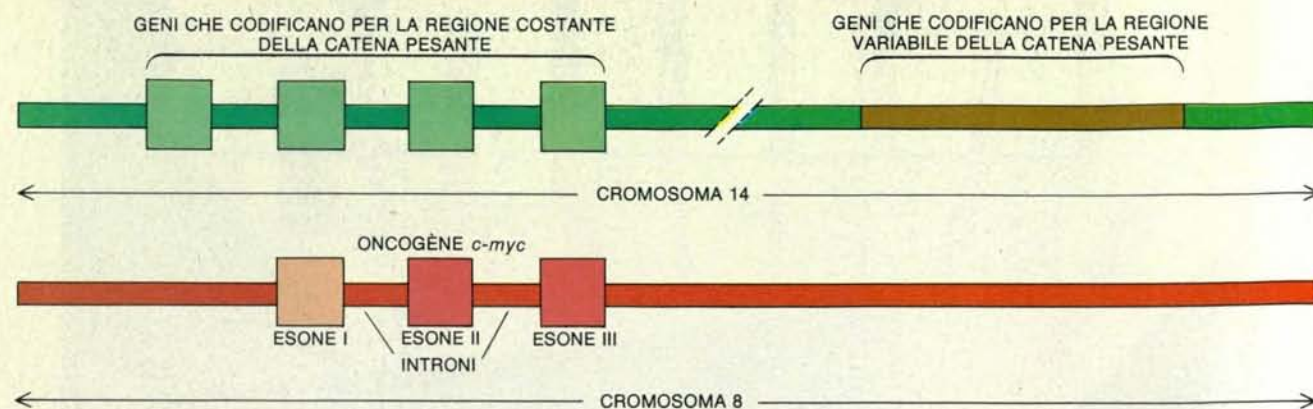
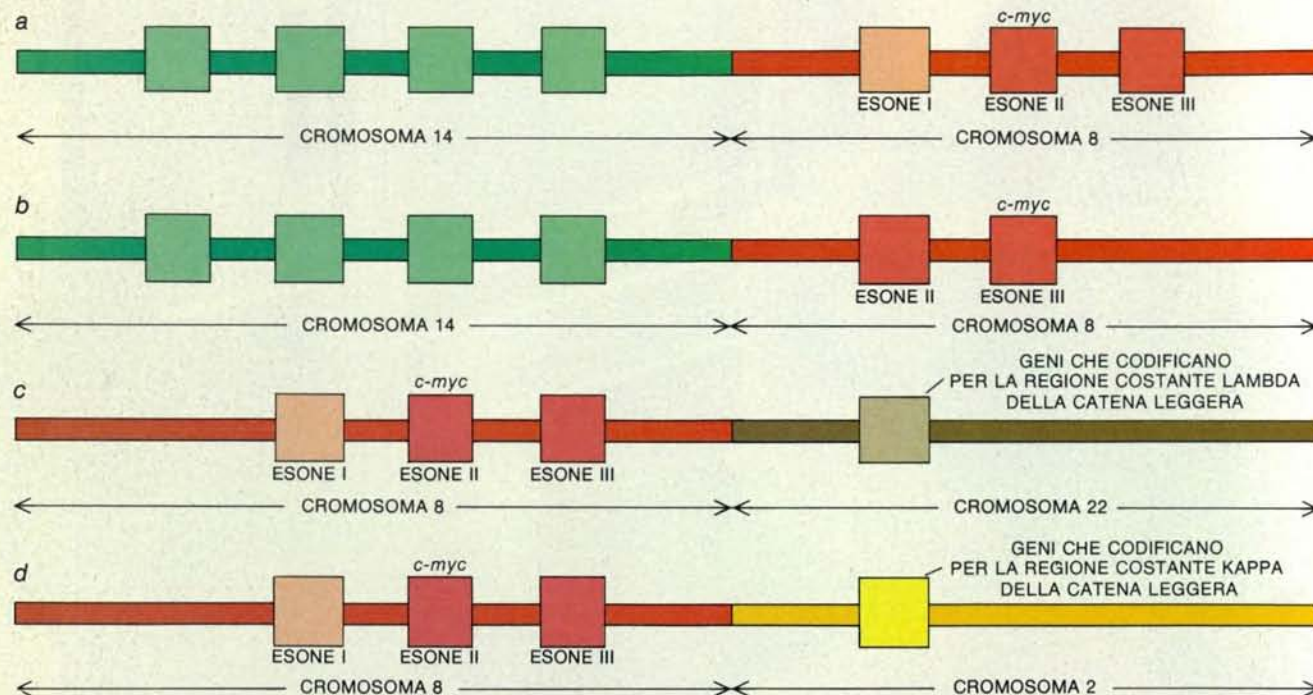
mu Tonegawa del Massachusetts Institute of Technology (MIT) hanno trovato che sequenze intensificatrici sono presenti nel segmento di DNA che codifica per un tipo di regione costante della catena pesante dell'immunoglobulina. Recenti studi compiuti al Wistar Institute suggeriscono l'ipotesi che sequenze intensificatrici supplementari siano presenti nel locus per la catena pesante. Inoltre, David Baltimore e collaboratori al MIT hanno trovato sequenze di questo genere nei segmenti cromosomici che codificano per la regione costante kappa della catena leggera.

Questi risultati suggeriscono un possibile meccanismo per il linfoma di Burkitt: le traslocazioni cromosomiche in una cellula *B* giustappongono l'oncogène *c-myc* alle sequenze intensificatrici; queste sono in grado di attivare la trascrizione su distanze considerevoli. Il gene *c-myc* si esprime quindi nello stesso modo in cui i geni per le immunoglobuline si esprimono in una cellula normale *B*. In un certo senso l'espressione dell'oncogène *c-myc* diventa parte della funzione specializzata della cellula.

Studi recenti indicano che questo meccanismo dell'oncogenesi può essere responsabile di molte altre forme maligne che interessano cellule del sistema immunitario umano. Jorge J. Yunis della Medical School dell'Università del Minnesota ha messo a punto un nuovo metodo di colorazione a bande dei cromosomi, che permette un riconoscimento altamente preciso di modificazioni cromosomiche specifiche in cellule maligne. Le traslocazioni tra cromosoma 14 e segmenti o del cromosoma 11 o del cromosoma 18 sono comuni nei linfomi delle cellule *B* degli adulti, nelle leucemie croniche umane da cellule *B* e nel mieloma multiplo. Questa osservazione, assieme alla conoscenza che abbiamo del ruolo del locus per la catena pesante dell'immunoglobulina sul cromosoma 14 nel linfoma di Burkitt, indica che gli oncogèni umani possono trovarsi sui cromosomi 11 e 18; questa congettura è sostenuta dai risultati del lavoro di Yoshishide Tsujimoto che lavora al Wistar Institute in collaborazione con Yunis e Nowell. Da parte nostra, abbiamo trovato che i punti di rottura in queste traslocazioni erano costantemente raggruppati in brevi segmenti sul cromosoma 11 o 18; inoltre essi si trovano sempre di fronte al segmento del cromosoma 14 che codifica per la regione costante della catena pesante. Per i due ipotetici oncogèni sui cromosomi 11 e 18 abbiamo proposto le designazioni *bcl-1* e *bcl-2* (linfoma delle cellule *B*/leucemia 1 e 2).

Osservazioni compiute nel corso dello studio del linfoma di Burkitt schiudono due nuove importanti aree alla ricerca. In primo luogo, vi è la questione delle sequenze intensificatrici. Quali sono le precise sequenze di DNA che hanno questa funzione e in che modo aumentano il livello di trascrizione di certi geni? L'altra area riguarda l'oncogène *c-myc*. Qual è la sua funzione in una cellula normale e perché l'espressione costitutiva di *c-myc* a livelli elevati dovrebbe provocare cancro?

Oltre a queste nuove aree di ricerca, il nostro lavoro suggerisce nuovi modi di affrontare sperimentalmente lo studio dei neoplasmi delle cellule *B*. Molte di queste forme maligne interessano la traslocazione di un oncogène ignoto nel locus per la catena pesante sul cromosoma 14. Questo locus è relativamente ben conosciuto e vi sono delle sonde costituite da acidi nucleici che permettono a un ricercatore di studiare segmenti di DNA vicino a esso. Dato che le traslocazioni tendono a portare l'oncogène in stretta prossimità del locus per la catena pesante, queste sonde offriranno ai ricercatori il mezzo per identificare, isolare e caratterizzare i geni connessi con la maggioranza delle neoplasie umane a carico delle cellule *B*. In questo modo, la conoscenza della genetica della produzione degli anticorpi permetterebbe di conoscere anche la struttura genetica degli oncogèni appena isolati.

La ricerca sulle traslocazioni cromosomiche può anche portare a nuovi metodi di diagnosi e di caratterizzazione di

Il linfoma di Burkitt può essere provocato da varie traslocazioni. Nel caso più comune (*a*) tutti e tre gli esoni (sequenze di DNA che codificano per le proteine) dell'oncogène *c-myc* si spostano dal cromosoma 8 a una sezione del cromosoma 14 adiacente ai geni che codificano per la regione costante della catena pesante dell'anticorpo. In alternativa (*b*), il cromosoma 8 si rompe nel primo introne (segmento di DNA «nonsenso», cioè non trascritto in RNA messaggero), nel qual caso solo due esoni si spostano sul cromosoma 14. In altre traslocazioni, il *c-myc* rimane sul cromosoma 8, mentre i geni che codificano per la regione costante della catena leggera dell'anticorpo si uniscono a esso. In uno di questi casi (*c*), i geni che codificano per le regioni costanti del tipo «lambda» sono traslocati dal cromosoma 22; i geni del cromosoma 2, che codificano per le regioni costanti «kappa», possono anch'essi prendere parte a questa traslocazione (*d*).

tumori maligni del sistema immunitario. I punti di rottura sui cromosomi, per esempio, si concentrano entro corti segmenti di DNA nei tumori maligni a carico di cellule *B* che mostrano traslocazioni tra i cromosomi 11 e 14, o tra i cromosomi 14 e 18. Dovrebbe pertanto essere possibile mettere a punto sonde di DNA specifiche per questi piccoli segmenti. Un campione di tessuto potrebbe allora essere prelevato dalla regione colpita e le sonde di DNA potrebbero essere utilizzate per determinare esattamente quale tipo di riassestamento cromosomico è responsabile del cancro.

Risultati conseguiti molto di recente indicano che quanto è stato appreso per le neoplasie a carico delle cellule *B* è applicabile anche alle neoplasie a carico delle cellule *T*, che sono l'altro principale componente del sistema immunitario. Uno di noi (Croce), in collaborazione con Rovera e con Mark M. Davis della Stanford University, ha trovato che il gene per la catena alfa dei recettori delle cellule *T* si trova nella regione del cromosoma 14 che è interessata in alcune traslocazioni caratteristiche di certe neoplasie che colpiscono le cellule *T*.

I nostri studi sul meccanismo che sta alla base del linfoma di Burkitt hanno così implicazioni che vanno al di là di questa malattia. Le traslocazioni nel linfoma di Burkitt forniscono in apparenza un modello per la maggioranza delle forme di cancro umano a carico delle cellule *B* (e forse anche di quelle a carico delle cellule *T*). Inoltre la conoscenza del meccanismo della traslocazione fornirà strumenti sperimentali potenti non solo per lo studio di altre forme di cancro, ma anche per lo studio dei meccanismi che controllano l'espressione genetica durante lo sviluppo e il funzionamento normali del sistema immunitario umano.

Una fortezza e un centro funerario neolitici

Scavi compiuti a Hambledon Hill, nell'Inghilterra sudoccidentale, rivelano che verso il 3600 a.C. vi si sviluppò un centro funerario, sul cui sito venne costruita in epoca successiva una grande fortezza

di R. J. Mercer

Hambledon Hill è un punto di riferimento di proporzioni imponenti per chi si trovi nella valle scavata dal fiume Stour attraverso il paesaggio calcareo dell'Inghilterra sudoccidentale. Un pastore del Neolitico che attorno al 3400 a.C. avesse rivolto lo sguardo verso la cima della collina avrebbe osservato uno spettacolo grandioso. Hambledon Hill era allora coronato da un'immensa cinta difensiva con tre terrapieni concentrici. Quello interno, il più imponente dei tre, era sostenuto da 10 000 travi di quercia del diametro di un palo telegrafico. Nel fossato attorno ai terrapieni crani umani collocati a un certo intervallo l'uno dall'altro conferivano una nota lugubre alle fortificazioni.

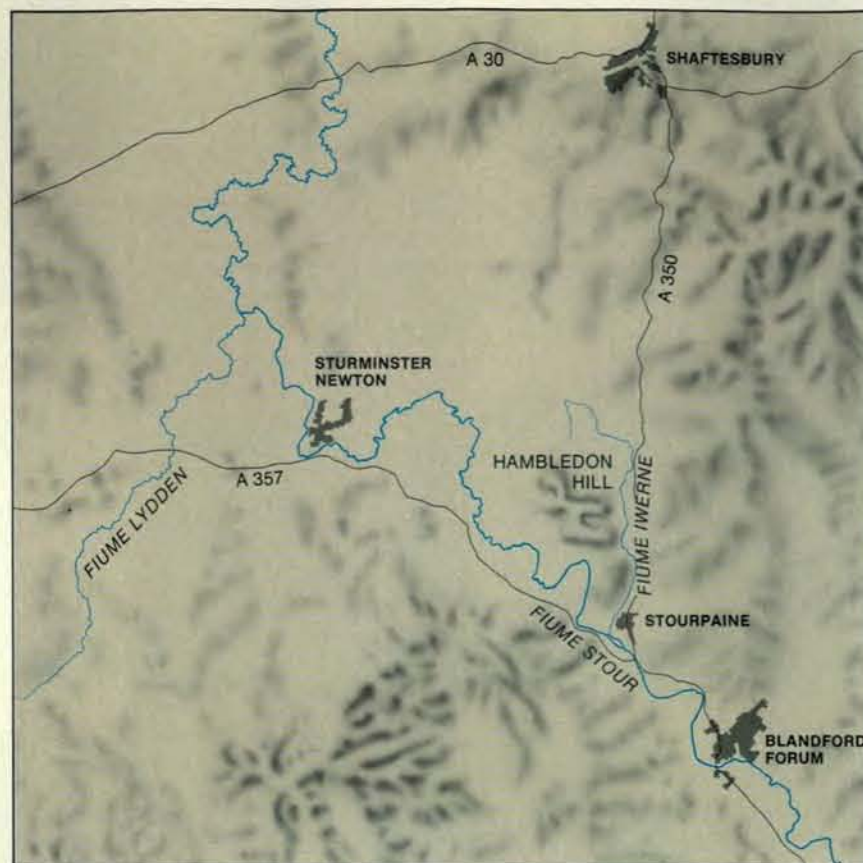
Il complesso neolitico di Hambledon Hill non aveva sempre avuto una funzione difensiva. L'imponente fortezza fu la fase finale in un processo di modificazione e di espansione che ebbe inizio attorno al 3600 a.C. e che potrebbe essere durato varie centinaia di anni. Dal punto di vista archeologico, più sorprendente delle fortificazioni è la possibilità che, ai suoi inizi, Hambledon Hill sia servito come scenario per complessi rituali funebri. Pare che, quando moriva un membro di una delle comunità vicine alla collina, la salma venisse esposta alle intemperie in un'area ben precisa. In alcuni casi accanto al corpo venivano forse deposte offerte di oggetti preziosi. Una volta che la carne si era staccata dalle ossa, alcuni cadaveri venivano probabilmente prescelti per essere sepolti in altra sede, con ulteriori cerimonie.

L'intenso programma edilizio di Hambledon Hill fu il risultato di importanti sviluppi sociali verificatisi nel Neolitico, che ebbero inizio in Gran Bretagna attorno al 4000 a.C. con il passaggio dalla caccia e raccolta all'agricoltura. Le prime comunità agricole non avevano probabilmente una base economica abbastanza stabile per costruire insediamenti permanenti; sono state trovate infatti ben poche tracce delle comunità agricole più antiche. A distanza di pochi secoli, però, la stabilità economica dovuta all'agricoltura permise a una parte dei membri della comunità di passare ad attività diverse dall'agricoltura stessa o dalla pastorizia. Le fortificazioni e i rituali funebri a Hambledon Hill furono una conseguenza dell'affrancamento di ingenti energie umane in seguito allo sviluppo dell'agricoltura, da attività di sussistenza. Hambledon Hill è fra i maggiori siti neolitici finora riportati alla luce in Europa ed è uno fra i pochi in cui fossero molto sviluppati i rituali funebri. Questo sito ci aiuta perciò a delineare un nuovo quadro della vita in Gran Bretagna durante il Neolitico.

In che modo è stato ricostruito il quadro della vita e dei rituali a Hambledon Hill? La campagna di scavi e rilievi sul campo, nota come Hambledon Hill Excavation and Field Survey Project, si protrasse dal 1974 al 1982 e rappresentò un'azione di emergenza per salvare opere dell'Età della pietra dall'invasione dell'agricoltura moderna. Da molto tempo si sapeva che a Hambledon Hill c'erano manufatti del Neolitico. Già nel 1913 l'architetto e antiquario inglese Heywood Sumner aveva scoperto dei resti e disegnato una fra le prime piante di un sito neolitico fortificato pubblicate in Gran Bretagna.

Una breve missione esplorativa compiuta nel 1959 da Desmond Bonney, della Royal Commission for Historical Monuments, dimostrò che sulla collina esistevano altri terrapieni risalenti al Neolitico ed espresse l'opinione che il complesso poteva meritare uno scavo su vasta scala. All'inizio degli anni sessanta si era cominciato ad arare il suolo per preparare pascoli per le pecore. Se quest'operazione fosse proseguita a lungo avrebbe finito col distruggere le prove necessarie per la comprensione di quel grande monumento. Pertanto, nel 1974 fu intrapreso uno scavo che, alla fine, interessò un'area di circa 60 000 metri quadrati.

Il lavoro sul campo portò alla luce sulla collina i resti di varie strutture neolitiche correlate *(si veda l'illustrazione a pagina 67)*. La maggior parte dei manufatti connessi con riti funebri è stata trovata all'interno di una grande area recintata (il recinto principale) con passaggi sopraelevati, o rialzi al centro della collina. Due lunghi tumuli, bassi monticelli allungati che potrebbero avere avuto una funzione cerimoniale, erano disposti l'uno di fronte all'altro ai due lati del recinto principale, uno a sud e l'altro a nord. Uno spazio recintato minore (il recinto di Stepleton) occupava la sommità del contrafforte sudorientale della

Hambledon Hill è un affioramento di chalk (un calcare biancastro) che domina un'area di ricchi pascoli. Durante l'Età della pietra la collina fu coronata da una grande necropoli e poi da una fortezza. Nella fotografia si può vedere ben poco dei monumenti neolitici. I terrapieni sul contrafforte settentrionale della collina *(in basso)* appartengono a un forte dell'Età del ferro, costruito su parte delle fortificazioni neolitiche. I resti connessi a riti funebri del Neolitico furono perloppiù riportati alla luce al centro della collina e sul contrafforte di Stepleton *(in alto a sinistra)*. Le fortificazioni del Neolitico proteggevano l'intera sommità della collina ed erano particolarmente imponenti sulle pendici meridionali e occidentali *(in alto a destra)*.

La posizione strategica di Hambledon Hill fu uno dei motivi della scelta di questa località per l'insediamento da parte di una comunità del Neolitico inferiore. La collina domina il corridoio scavato dal fiume Stour nel paesaggio calcareo dell'Inghilterra sudoccidentale e sovrasta i ricchi pascoli della valle di Blackmore verso ovest. Attorno al 4000 a.C., all'inizio del Neolitico, in quest'area potrebbe essersi concentrata una popolazione attratta dalla disponibilità di due risorse chiave: ricchi pascoli e selce per la produzione di utensili. La fortezza neolitica in cima alla collina doveva essere visibile a chilometri di distanza ai membri delle comunità che vivevano nelle pianure ai piedi della collina.

collina. Esso fu probabilmente un'area residenziale per almeno una parte del tempo in cui fu usato nel Neolitico.

Le fortificazioni difensive che cingevano la cima della collina erano particolarmente solide sulle pendici meridionale e occidentale, fra il recinto di Stepleton e il recinto principale. Può darsi che ci fosse anche un terzo recinto in una posizione strategica per la difesa sul contrafforte settentrionale della collina, ma in tal caso si troverebbe sotto un forte collinare molto posteriore, risalente all'Età del ferro, e non è stato ancora scavato. Malgrado la notevole vicinanza, le strutture difensive e i monumenti funerari non furono con ogni probabilità contemporanei; determinare la cronologia delle fasi di costruzione sul sito è stata una delle sfide più grandi del progetto di ricerca sul campo.

Oltre alla posizione strategica, senza dubbio vari fattori economici influirono sulla scelta dell'imponente collina come sito di costruzione. La collina domina un'area ben fornita di due risorse d'importanza critica all'inizio del Neolitico: terra da pascolo e selce. I ricchi pascoli ai piedi della collina comprendono la Valle di Blackmore a ovest e i fianchi degli altipiani calcarei di Cranborne Chase a est. I terreni calcarei, di quel calcare biancastro di origine pelagica chiamato *chalk*, sono ricchi di selce e vi abbondano le armi da caccia e le asce in selce dei cacciatori preistorici.

L'area attorno a Hambledon Hill, essendo una regione fertile in cui erano già concentrati cacciatori preistorici, fu un sito naturale per lo sviluppo dell'agricoltura già in epoca molto antica. Il processo di sviluppo potrebbe essere stato accelerato dall'arrivo di immigranti dal continente europeo, in possesso di tecniche e materiali nuovi, ma questa parte del Neolitico in Gran Bretagna non è ancora ben compresa. In ogni caso attorno al 4000 a.C. era già in corso la transizione a un'economia agricola.

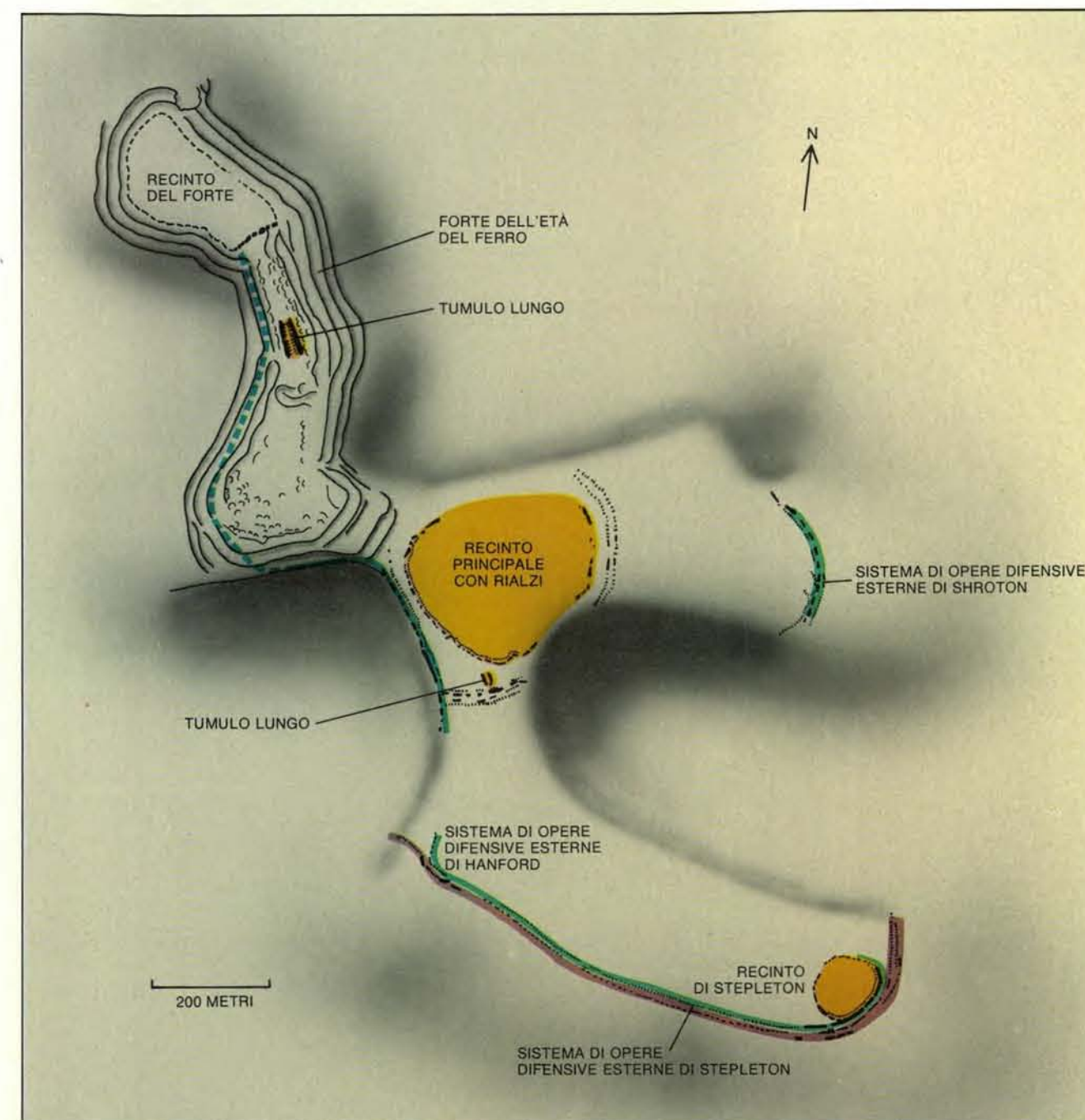
Nel passaggio da gruppi dediti alla caccia e raccolta a comunità agricole stabili ebbe un ruolo importante un nuovo tipo di struttura architettonica: il recinto circondato da mura. Questo favoriva il controllo delle risorse, delimitando aree in cui venivano svolte attività specializzate (come la fabbricazione di utensili) e fornendo alla comunità una difesa da attacchi esterni. Nell'Inghilterra meridionale sono note una sessantina di aree cintate neolitiche, di superficie compresa fra uno e 70 ettari. In generale attorno alla circonferenza esterna del muro di cinta correva un fossato. La struttura complessiva variava da una semplice area cinta da un solo fossato a siti anche con cinque anelli concentrici di fossati.

Facendo astrazione dalla complessità della loro struttura, quasi tutte le aree recintate neolitiche condividevano una caratteristica costruttiva: i fossati attorno all'area cintata non erano continui, bensì interrotti a intervalli irregolari da rialzi perpendicolari all'asse maggiore della depressione circolare. La presenza di questi passaggi sopraelevati induce a pensare che i fossati non fossero concepiti tanto come barriere difensive quanto piuttosto come cave da cui estrarre i materiali utilizzati nella costruzione dell'argine o terrapieno interno. In alcuni casi il terrapieno si è conservato, anche se ovviamente le sue dimensioni sono state molto ridotte dall'erosione. Materiale da costruzione e stile del terrapieno dipendevano probabilmente dalle risorse disponibili e dalla funzione dell'area recintata.

A Hambledon Hill il recinto principale, al centro della collina, formò uno dei siti di maggiore interesse per le ricerche sul campo. Poiché nel 1974, all'inizio del progetto di scavo e ispezione del sito, si sapeva molto poco su quest'area monumentale, si adottò in principio una strategia molto semplice: si scavò il 20 per cento circa dell'area all'interno del recinto principale, nel desiderio di accertare a quale uso esso fosse stato adibito in tempi neolitici. Fu scavata inoltre press'a poco la stessa percentuale di fossato esterno e questi lavori fornirono ulteriori informazioni sulla funzione del sito. Lo scavo del fossato permise di fissare una cronologia delle fasi dell'uso di questa parte del complesso di Hambledon Hill, cronologia che si fondò su datazioni con il carbonio radioattivo e su una ricostruzione accurata degli strati accumulatisi nel fossato a partire dall'epoca in cui fu scavato.

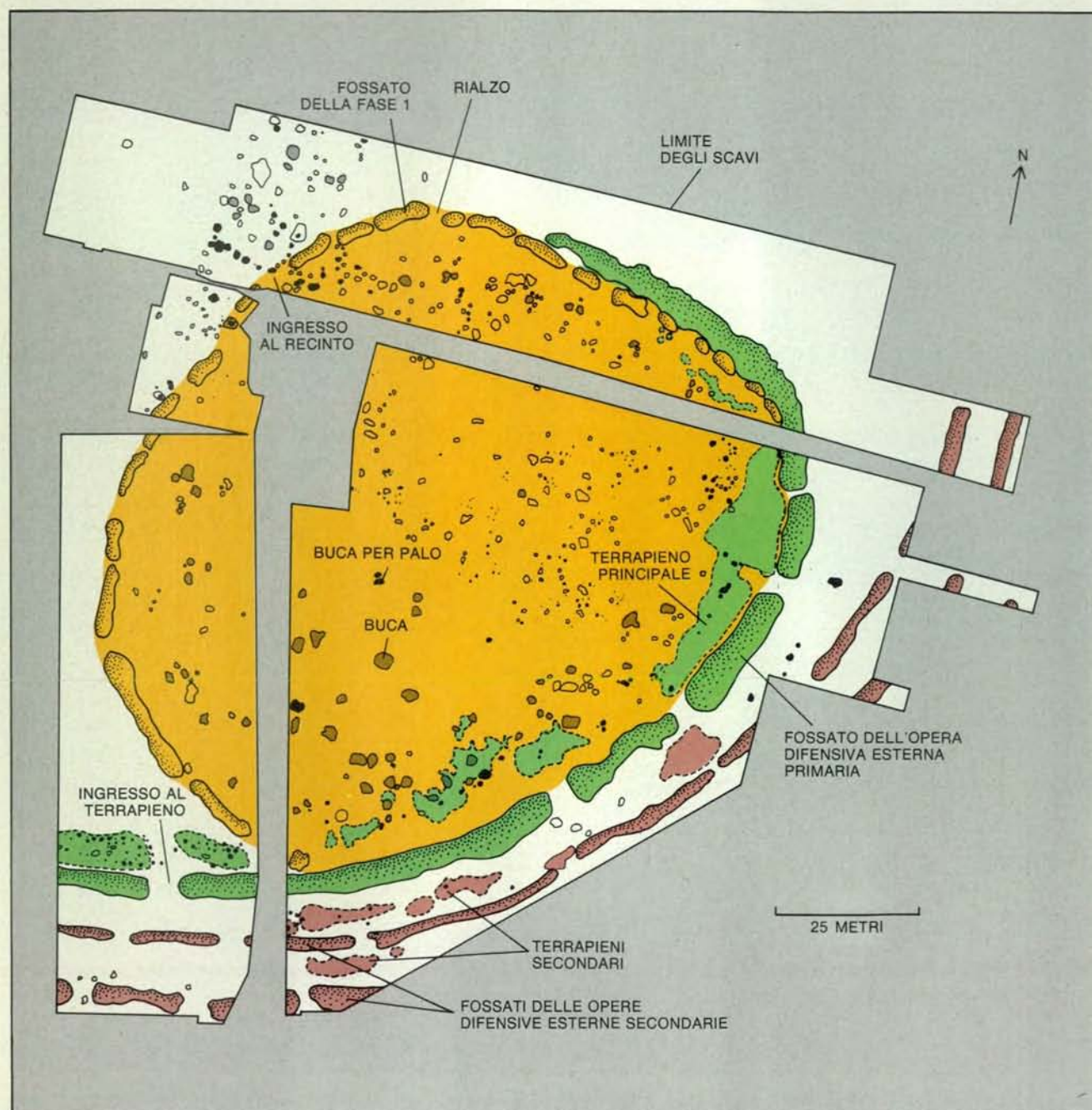
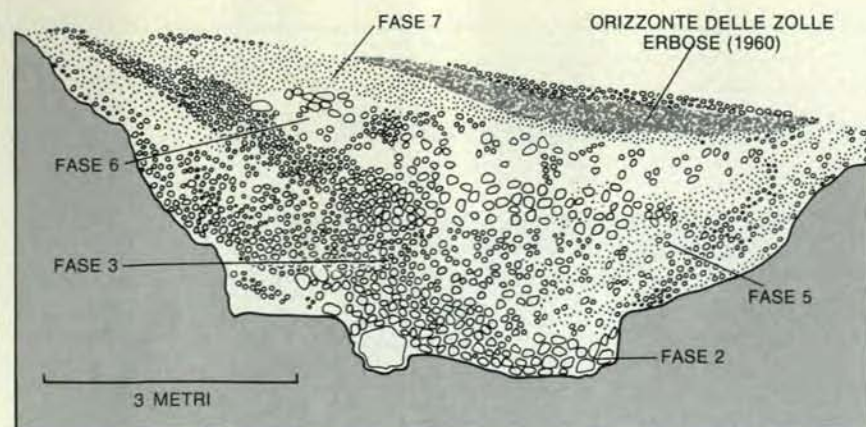
Le prime operazioni rivelarono che il sottosuolo di Hambledon Hill aveva sofferto gravi danni per l'erosione e i lavori agricoli. Fu subito evidente che la collina era stata arata verso la fine del Neolitico, nell'Età del bronzo e del ferro, oltre che in epoca romana e medioevale. I millenni di aratura avevano asportato da 70 centimetri a un metro della parte superiore del suolo.

La rimozione di una quantità così considerevole di suolo pose un grave problema archeologico, perché assieme al suolo erano andate perdute in gran parte, dalle fondamenta degli edifici, le buche per pali e altre strutture in legno. Oggi perciò siamo in grado di dire ben poco sulle abitazioni e su altre strutture che potrebbero essere esistite a Hambledon Hill. Questa è una perdita molto grave. Molte fra le buche nel recinto principale erano però così profonde che la loro parte inferiore poté sopravvivere all'aratura: esse hanno fornito così molti particolari affascinanti su pratiche rituali del Neolitico.

Era chiaro che in molti casi le buche erano state scavate e lasciate aperte, cosicché un sedimento di chalk naturalmente eroso aveva finito col depositarsi sul loro fondo. Solo quando ciò era accaduto, erano stati deposti nella buca oggetti scelti con cura. Questi gruppi di oggetti, comprendenti vasellame, asce in pietra e corna di cervo nobile, erano presumibilmente offerte rituali che forse accompagnavano le spoglie esposte nel recinto.

La pianta del sito illustra in che modo si sviluppò Hambledon Hill. Nella sua fase iniziale *(in giallo)* il sito fu un centro per cerimonie funebri. Il recinto principale con passaggi sopraelevati, o rialzi, al centro della collina, era il luogo in cui venivano inizialmente esposte le salme. Dopo che la carne si era staccata dalle ossa, i resti di alcune di esse probabilmente venivano sepolti in un paio di tumuli lunghi situati a nord e a sud del recinto principale. Il recinto di Stepleton potrebbe essere stato un luogo residenziale per un piccolo gruppo privilegiato di persone che guidavano le cerimonie funebri. Quando il centro funerario declinò, la comunità neolitica trasformò il sito in fortezza. Un terrapieno fu costruito sulle pendici meridionale e occidentale e sul contrafforte di Shroton *(in verde)*. Un terzo recinto, sul contrafforte settentrionale, sotto il forte dell'Età del ferro, potrebbe essere stato il centro di comando della fortezza. Altri due terrapieni *(in color prugna)* rafforzarono in seguito quello principale lungo la pendice meridionale, che degrada dolcemente e quindi è maggiormente vulnerabile.

Gli oggetti trovati avevano senza dubbio un valore di alto prestigio per i membri della comunità neolitica. I frammenti di vasellame recuperati dalle buche sono chiaramente i resti di vasi completi. Rispetto ai cocci rinvenuti in altre parti del sito, essi comprendono un'alta percentuale di oggetti importati da regioni lontane, come la Cornovaglia e il Devon. In parte i recipienti sono molto grandi e furono eseguiti con un'abilità che pungolò altri artigiani all'imitazione (riuscita solo in parte) con materiali locali.

Fra altri oggetti importanti rinvenuti nelle buche vi sono asce di pietra. L'analisi del materiale con cui furono fabbricate dimostra che esse provenivano dalla Cornovaglia, dal Galles meridionale e da Borrowdale nel Cumberland. Il pregio di questi depositi funerari per gli abitanti del Neolitico è suggerito dal fatto che le asce erano oggetti abbastanza ambiti da giustificarne l'importazione da distanze che per quei tempi erano enormi: ben 400 chilometri nel caso di Borrowdale. Ancora più sorprendente fu il ritrovamento di due asce, una di nefrite e una di giadeite, che non sono originarie della Gran Bretagna ma che verosimilmente provengono dalla Bretagna o forse da ancor più lontano.

Gli oggetti trovati nel recinto principale suggerirebbero perciò una connes-

Il recinto di Stepleton era probabilmente una zona residenziale *(in giallo)*, come indicano i rifiuti che vi sono stati trovati. Lo scavo del fossato appartenente alla fase 1 della storia del sito ha fornito materiale per un argine di cui rimane ben poco. In alto a sinistra vi era l'ingresso al recinto. Le buche per pali all'interno del recinto stesso potrebbero essere state fondazioni per case d'abitazione; le buche contenevano i resti di banchetti e della lavorazione della selce. Dopo il crollo dell'argine originario, su parte del recinto furono costruite opere di fortificazione. Il fossato delle opere di difesa esterne primarie ha fornito materiale per il terrapieno principale *(in verde)*, a cui si accedeva attraverso l'ingresso per la porta in basso a sinistra. Dai fossati secondari deriva, invece, il materiale per i terrapieni più piccoli *(in color prugna)*.

Nella sezione trasversale del fossato attorno al recinto principale si osservano strati che hanno contribuito a determinare la cronologia degli eventi svoltisi nel sito durante il Neolitico. Nella fase 1 si ebbe lo scavo del fossato stesso, cosicché in esso non compaiono resti coevi. Il chalk estratto fu usato per costruire il muro del recinto. Fra i depositi della fase 2 furono trovate ossa umane e quelle che sembrerebbero offerte rituali, costituite da oggetti in corno di cervo nobile, suppellettili in pietra e vasellame. Lo strato della fase 3 è costituito dal muro del recinto principale, crollato nel fossato all'epoca in cui il recinto non fu più usato. Lo strato della fase 5 riflette il nuovo scavo del fossato effettuato con intenti ritualistici e la deposizione di nuove offerte. Nella fase 6 sul fossato originario fu costruito un cumulo (*cairn*) di selce. La fase 7 mostra un accumulo di suolo agricolo posteriore al Neolitico. (Nel punto in cui fu eseguita questa sezione non erano presenti depositi della fase 4, che assomigliano a quelli della fase 5.)

sione con pratiche rituali. Lavori di scavo nel fossato attorno al recinto principale sono valsi a confermare che venivano eseguiti effettivamente dei rituali e hanno anche fornito informazioni sulla loro natura. Il fossato fu scavato in origine per ricavarne materiali per la costruzione del terrapieno attorno all'area recintata. I resti del terrapieno conservano ben poco della passata grandiosità e inducono a pensare che esso fosse formato da una struttura di sostegno in legno, una sorta di lungo cassone, riempito con una massa di chalk tratto dal fossato che dava luogo a una barriera imponente ma in definitiva instabile.

Una volta ultimato il terrapieno, sul fondo del fossato, a mano a mano che materiali dilavati dai lati della depressione venivano trasportati in basso, andò accumulandosi un limo pastoso. In molti punti lungo la circonferenza del fossato il limo che si era ammassato sul fondo sembra esserne stato asportato con cura. Il materiale raccolto in questo modo fu probabilmente usato per la manutenzione e la riparazione del terrapieno.

La rimozione del limo primario da parte dei costruttori ci ha privati di informazioni archeologiche sull'uso iniziale dell'area recintata. Dopo l'asportazione del limo, però, fu deposta sul fondo del fossato una raccolta di oggetti di carattere verosimilmente rituale. I depositi, che erano probabilmente offerte e che, in origine, potrebbero essere stati contenuti in sacchi di pelle, comprendono ossa umane, ossa animali, utensili in selce e vasellame.

Oltre a questi oggetti, sul fondo del fossato ne furono collocati altri, connessi in modo più chiaro con l'inumazione di salme. Una serie di crani fu disposta a intervalli irregolari, con il lato destro rivolto verso l'alto. Negli strati che costituiscono il fondo del fossato è dispersa anche una considerevole quantità di ossa umane infrante. Fra questa massa di ossa, sopravvivono intatte due sepolture di bambini sotto cumuli di selce ben costruiti.

Forse più significativi sono il tronco e i femori di un giovane di circa 15 anni la cui salma era rimasta chiaramente esposta per qualche tempo nel fossato in uno stato piuttosto avanzato di decomposizione. Quando essa aveva cominciato a smembrarsi, alcune sue parti erano forse state trascinate sul fondo da cani o altri predatori, che ne rosicchiarono estesamente le ossa.

L'ipotesi che, a Hambledon Hill, i corpi dei defunti venissero esposti intenzionalmente nel recinto principale o nei pressi potrebbe aiutare a rispondere a due domande importanti sulle pratiche funerarie del Neolitico. Lo scavo di tumuli lunghi in altri siti ha dimostrato che in essi venivano definitivamente deposti scheletri o parti di scheletri che erano stati lasciati esposti altrove alle intemperie. Ma dove aveva avuto luogo questa esposizione?

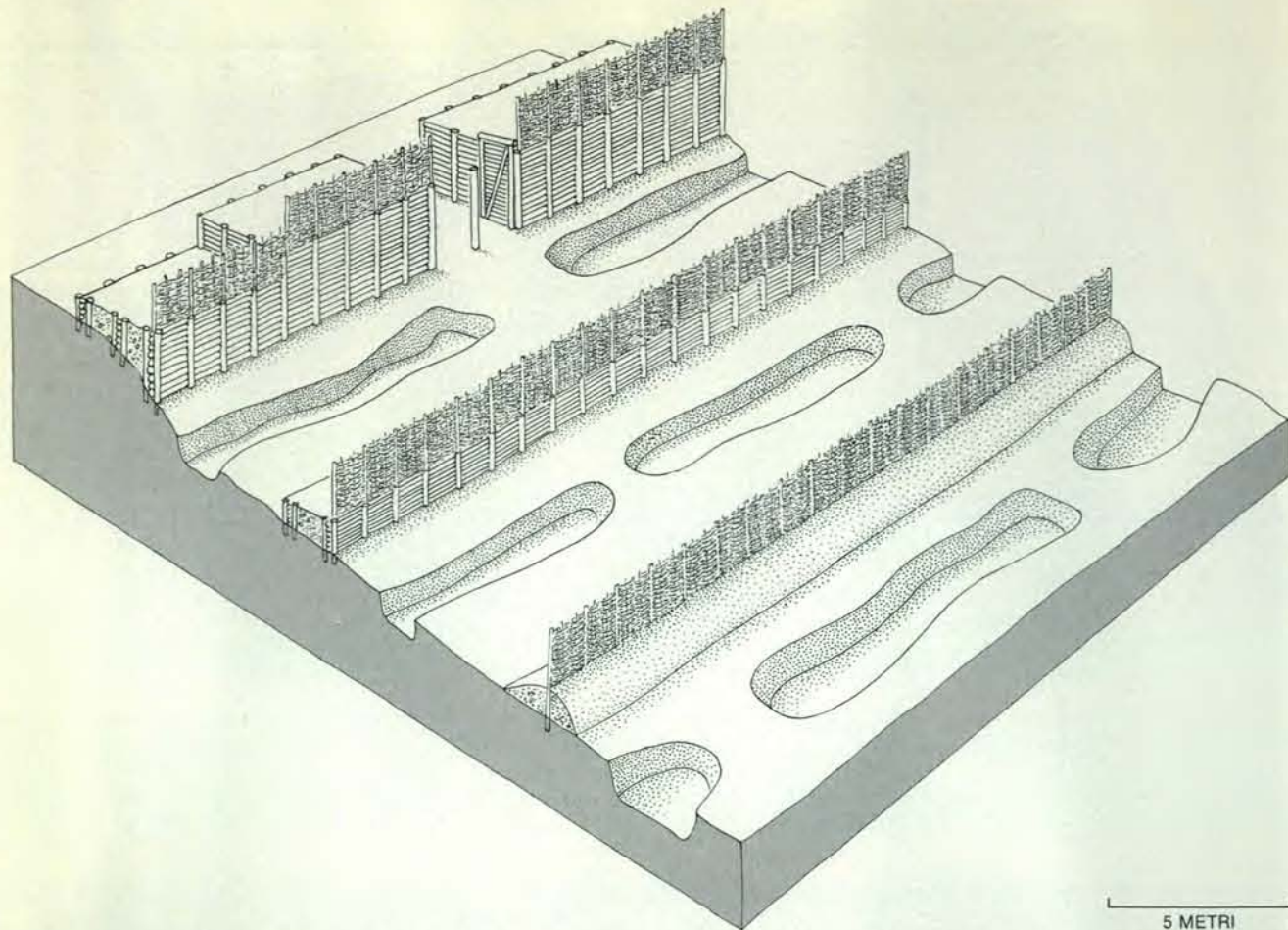
Nelle ossa recuperate dai tumuli si riscontra, inoltre, un curioso squilibrio. In generale, gli scheletri o parti di scheletri comprendono i resti di un numero relativamente piccolo di donne e bambini. L'assenza di ossa di bambini è particolarmente strana se si considera il tasso di mortalità senza dubbio molto elevato fra neonati e bambini in epoca preistorica. Che cosa ne fu allora delle ossa delle donne e dei bambini?

Un'ipotesi ragionevole suggerita dalle nostre scoperte è che i corpi dei membri della comunità venissero esposti in centri cerimoniali come il recinto principale di Hambledon Hill. Dopo che la carne si era distaccata dalle ossa, alcuni scheletri venivano probabilmente prescelti per l'inumazione nei tumuli lunghi. Se l'ipotesi è esatta, i centri di esposizione, come il recinto principale di Hambledon Hill, rientravano in un rituale comprendente anche la sepoltura nei tumuli lunghi.

Quest'ipotesi è corroborata in qualche misura dal fatto che il 60 per cento della grande quantità di ossa trovate nel recinto principale di Hambledon Hill proviene da bambini in tenera età. Pare, inoltre, che le ossa di individui di sesso maschile e di sesso femminile siano presenti in proporzioni press'a poco uguali. Tutti i membri della comunità o di un sottogruppo della comunità risultano dunque rappresentati. Per istituire una connessione più diretta fra il recinto principale e i tumuli sarebbero però necessarie prove tratte dai tumuli lunghi presenti nel sito.

Prove dirette di questo genere non sono ancora disponibili per Hambledon Hill. Il tumulo a sud, lungo 20 metri, fu distrutto dai bulldozer durante i lavori eseguiti per migliorare i pascoli negli anni sessanta. I bulldozer ne cancellarono il rialzo, distruggendo gran parte degli elementi che sarebbero occorsi per verificare l'ipotesi. Le somiglianze fra i manufatti rinvenuti nella fossa del tumulo e quelli venuti alla luce nel fossato del recinto principale inducono a pensare che, nella mente dei loro costruttori, i due monumenti fossero connessi. La perdita del rialzo del tumulo ci toglie però la possibilità di istituire una connessione fra i corpi esposti nel recinto principale e quelli sepolti nel tumulo.

Il rialzo del tumulo a nord, che è lungo 66 metri, è intatto, ma il tumulo non è stato ancora scavato. Ciò è dovuto al fatto che il tumulo a nord si trova in una posizione protetta all'interno del forte posteriore, che risale all'Età del ferro. Poiché il progetto per Hambledon Hill fu, in un certo senso, un'opera di salvataggio, i lavori si concentrarono sui monumenti più vulnerabili, rimandando lo scavo del tumulo settentrionale, che non correva alcun rischio. Il materiale trovato nel recinto principale di Hambledon Hill è certo stimolante, ma una soluzione definitiva del problema della connessione fra tale recinto e i tumuli lunghi richiederà altre ricerche.

Nella parte finale della storia del complesso funerario sulla cima della collina assunse importanza dominante il recinto di Stepleton. Molto più piccolo del recinto principale, esso fu riconosciuto come una struttura del Neolitico per mezzo di ispezioni *in situ*, di prospezioni del suolo e di fotografie aeree. È probabile che nelle fasi iniziali della costruzione di Hambledon Hill il recinto di Stepleton fosse una struttura piccola, semplice, con un ingresso che guardava a monte, verso il recinto principale.

Tre terrapieni costeggiavano il fianco meridionale di Hambledon Hill quando la cima della collina era occupata da una fortezza al culmine del suo splendore. Il terrapieno principale, quello più interno, era costituito da una massa di calcare racchiusa in un'incastellatura di legno formata da 10 000 travi di quercia grosse come pali telegrafici. A ogni ingresso erano presenti due porte che si chiudevano battendo su un grande palo centrale di quercia. Il riparo sopra i terrapieni era formato da piante giovani intrecciate. Anche il secondo terrapieno aveva una struttura a travi di quercia; il terzo era formato semplicemente da un accumulo di pietrisco calcareo. I fossati erano interrotti da passaggi sopraelevati, il che fa pensare che non servissero per la difesa, ma per fornire materiale per la costruzione dei terrapieni.

Le buche per pali trovate nel recinto di Stepleton dimostrano che in quel sito sorgevano costruzioni. I reperti sono troppo esigui per consentirci di ricostruire tali strutture, ma il materiale trovato nei pressi dell'area induce a pensare che dovessero essere abitazioni. Per esempio, i depositi più antichi nel fossato attorno al recinto contengono poche ossa umane (e nessun cranio), ma grandi quantità di rifiuti connessi alla lavorazione della selce e a quella delle corna di cervo nobile, che nel Neolitico erano tipiche attività domestiche.

Il recinto di Stepleton non era solo un luogo di lavoro: ossa di animali trovate in e attorno a esso indicano che vi avvenivano molti festini. Lo stato delle ossa dimostra che in questi banchetti si sprecava molta carne e ci sono ben pochi indizi del fatto che le ossa venissero frantumate per farne brodo o stufati di qualità scadente. Gli abitanti di Stepleton, quale che fosse la loro posizione sociale, apprezzavano il tipo di cibo più raffinato del Neolitico - la carne con osso arrosto - e non esitavano a gettar via i tagli meno pregiati. Il tipo dei consumi accertati all'interno del recinto e attorno a esso induce a pensare che a Stepleton possa esser vissuto un gruppo piccolo e piuttosto privilegiato. È forte la tentazione di supporre che esso presiedesse ai complessi rituali funebri che potrebbero aver connesso il recinto principale con i due tumuli lunghi.

Mentre Hambledon Hill era ancora al culmine delle sue fortune come centro funerario, ebbe inizio un periodo di trasformazioni che durò probabilmente due o tre secoli. L'effetto finale fu quello di convertire Hambledon Hill in una grande fortezza, ma il processo fu graduale e rituali di venerazione dei defunti furono certamente eseguiti ancora per molto tempo nel recinto principale.

La prima modificazione difensiva fu intrapresa probabilmente dopo che il recinto principale, il recinto di Stepleton e i tumuli avevano subito già un certo deterioramento. Lungo la parte meridionale della collina fu costruito un importante fossato con rialzi che era appoggiato a un terrapieno rafforzato da strutture lignee; la costruzione potrebbe aver racchiuso tutti i punti vulnerabili di avvicinamento alla sommità della collina, che ha un'area di circa 60 ettari.

Ignoriamo quasi tutto del primo sistema di fortificazioni che cingeva il sito perché esso fu distrutto poco tempo dopo essere stato costruito, durante le operazioni di ricostruzione e di rafforzamento delle difese. I vari segmenti del fossato furono approfonditi e un nuovo e più grande terrapieno rinforzato da strutture in legno fu costruito sui resti del precedente. Anche questa seconda opera avanzata di fortificazione proteggeva i 60 ettari della sommità della collina.

Come il terrapieno precedente, anche il secondo era irrobustito da una struttura di sostegno in legno a forma di lungo cassone. Puntelli verticali costituiti da travi di quercia grossi come pali telegrafici furono messi in opera a un intervallo di un metro circa l'uno dall'altro lungo la faccia esterna e quella interna del vallo. Per fornire stabilità alla struttura, i sostegni furono probabilmente collegati fra loro per mezzo di travi orizzontali che passavano attraverso il terrapieno stesso. Complessivamente, nella costruzione della struttura in legno potrebbero essere state usate circa 10 000 travi di quercia. Un progetto di tali dimensioni deve aver chiesto molto alle capacità di un'antica comunità agricola in termini di investimento di risorse di mano d'opera.

Mentre la sommità della collina veniva trasformata in una fortezza, il centro funerario iniziò a essere trascurato. In alcuni punti il terrapieno del recinto principale e quello del recinto di Stepleton rovinarono nei rispettivi fossati, e lo stesso si verificò anche per il più piccolo dei tumuli lunghi.

Nonostante le loro cattive condizioni, il recinto principale e i tumuli conservarono chiaramente qualcosa della loro funzione rituale originaria. L'esame degli strati nel fossato del recinto principale ha dimostrato che, dopo il crollo del terrapieno che circondava quest'area recintata, nel pietrisco calcareo che colmava il fossato furono scavate delle buche. Riempite di cenere, vasellame, ossa umane e animali, esse erano in qualche caso abbastanza profonde da giungere alla base del fossato originario.

In seguito, attorno alla circonferenza del vecchio fossato, fu scavato un solco stretto. In alcune parti del fossato, questo solco fu scavato e riscavato almeno quattro volte. Alcune sezioni del nuovo solco furono riempite con pietrisco calcareo, mentre altre furono lasciate scoperte, in modo che vi si accumulasse del fango; altre ancora furono colmate da ricchi depositi di ossa animali, vasellame e utensili di selce. Lo scavo delle buche e quello del solco potrebbero essere stati atti di venerazione legati ai precedenti riti funebri.

L'ultima fase dell'uso del recinto principale come centro di cerimonie fu la costruzione di un cumulo (*cairn*) lineare di selce sopra il fossato originario. Gli scavi eseguiti nel fossato attorno al più piccolo dei tumuli lunghi mostrano che anche qui prevalse la stessa sequenza di atti rituali: nel fossato riempito di pietrisco calcareo fu scavato con cura un solco e in seguito sul sito del vecchio fossato fu costruito un cumulo di selce. Questo parallelismo induce a pensare che nella mente dei costruttori i due monumenti fossero ancora connessi tra loro quali parti di un complesso funerario, benché di importanza ormai ridotta. Accanto al recinto di Stepleton, che - a quanto pare - rimase un'area residenziale, non si è trovata traccia di tali celebrazioni.

Mentre il carattere dei monumenti all'interno del terrapieno andava lentamente mutando, proseguiva il rafforzamento delle opere difensive. Dopo la costruzione dell'opera esterna principale, la costruzione di altri due terrapieni consolidò il versante meridionale, più vulnerabile, della collina. Il muro più esterno era formato da terra non sostenuta da strutture in legno; il terrapieno di mezzo era sostenuto da un'incastellatura di legno, a cassone, simile a quella che consolidava il muro interno.

I membri della comunità entravano nella fortezza sulla sommità della collina attraverso tre grandi ingressi, chiusi da portoni di legno. Un ingresso si trovava in prossimità del recinto di Stepleton; il secondo era sul contrafforte di Hanford, fra il recinto di Stepleton e il recinto principale; il terzo sul contrafforte orientale della collina. Ciascun portone comprendeva due battenti, ai lati di un grande palo centrale. Una carreggiata rivestita da grandi travi di quercia larga due metri e mezzo passava attraverso l'ingresso che si restringeva leggermente verso l'interno del terrapieno.

Quando la costruzione dei terrapieni esterni fu conclusa, la fortezza doveva essere visibile a vari chilometri di distanza dalla pianura a ovest, nella valle di Blackmore, dove pascolavano le greggi della comunità. I fianchi meridionale e occidentale della collina erano costeggiati da un terrapieno rinforzato da strutture in legno, lungo due chilometri e mezzo. Sul ripido fianco occidentale era stata scavata una terrazza per fornire una base stabile al terrapieno. Sul versante meridionale un'opera difensiva esterna, composta da vari fossati e lunga 1200 metri, costituiva un imponente ostacolo per eventuali aggressori.

Centro di comando delle massicce e impressionanti fortificazioni potrebbe essere stato il recinto sul contrafforte settentrionale della collina. Questo recinto, che è oggi quasi totalmente cancellato dal posteriore forte dell'Età del ferro, fu scoperto durante gli scavi grazie a fotografie aeree combinate ad assidue ispezioni sul terreno compiute da Roger Palmer, il topografo del progetto. Il recinto di quattro ettari e mezzo, che si trova in una posizione eccellente per la difesa, non è stato ancora scavato.

Il completamento delle tre cinte di terrapieni ha fatto di Hambledon Hill una grandiosa struttura difensiva. Eppure la fortificazione era tutt'altro che inespugnabile e, in effetti, la documentazione archeologica mostra che, dopo un attacco a cui seguì l'incendio di un lungo tratto dell'opera difensiva esterna sostenuta da strutture in legno, il sito venne abbandonato. Questo segmento del muro, lungo circa 200 metri, si trova sul contrafforte sudorientale della collina.

Difficilmente l'incendio fu accidentale, dal momento che l'intera struttura lignea fu consumata dal fuoco e i pali di quercia bruciarono fin nella buca che li conteneva. Per ottenere questo risultato, si deve pensare che il muro sia stato dato deliberatamente alle fiamme con torce. Consumate dal fuoco, le strutture esterne, le travi e poi il materiale di riempimento del terrapieno crollarono nel fossato. Gran parte di questo materiale è a sua volta bruciacchiato, attestando l'intensità dell'incendio.

Tanto gli attaccanti quanto i difensori subirono probabilmente delle perdite nella lotta attorno alla fortificazione in fiamme. Nel pietrisco furono trovati gli scheletri intatti di due giovani di sesso maschile, le cui condizioni testimoniano che furono rapidamente sepolti. Uno dei due stava probabilmente portando un bambino in tenera età, che rimase schiacciato sotto il suo peso quando egli cadde. Il giovane fu ucciso a quanto pare da una punta di freccia in forma di foglia finemente lavorata, che gli penetrò nella cavità toracica.

Un altro giovane di sesso maschile, abbandonato morto sul ciglio del fossato davanti all'opera difensiva esterna, non fu ricoperto dai detriti del terrapieno. Le condizioni delle ossa indicano che il suo corpo fu ben presto scoperto da predatori di ogni sorta. Un quarto scheletro, trovato nella parte superiore del materiale di riempimento del fossato del recinto di Stepleton, potrebbe esservi stato trascinato e smembrato da cani o lupi.

Alcune fra le vittime dell'attacco furono inumate con maggiori cerimonie. Sul lato nord del recinto di Stepleton sono state individuate due sepolture appositamente allestite, che potrebbero essere connesse all'incendio. Un corpo, quello di un giovane di sesso maschile, fu deposto con cura in una fossa che venne poi riempita con materiale calcareo recante chiare tracce di fuoco; l'unica fonte che si conosca di questo calcare è il terrapieno distrutto nell'attacco.

Le prove di una fine violenta del complesso sulla sommità della collina sono tutt'altro che ambigue, ma rimangono oscure le circostanze storiche dell'attacco. Attorno al 3300 a.C., durante il Neolitico medio, potrebbe esserci stato un periodo di disordini sociali, causati da fattori economici o ambientali. Vari siti del Neolitico furono abbandonati attorno a quest'epoca e alcuni sono cosparsi di punte di frecce in forma di foglia come quella trovata a Hambledon Hill.

Ma anche interrogativi più particolari, come quello riguardante la strategia usata nell'attacco finale a Hambledon Hill sono oggi senza risposta. Pare probabile che l'obiettivo centrale dell'attacco fosse il recinto fortificato sul contrafforte settentrionale della collina, su cui nell'Età del ferro fu costruito un forte. Gli scheletri e i materiali bruciacchiati trovati finora sono forse i resti di uno scontro preliminare.

Per dire se sia stato veramente così occorrerà eseguire scavi sul sito del forte sepolto sul contrafforte settentrionale della collina. Fortunatamente questo è stato acquistato di recente dal British Nature Conservancy Council e resterà perciò protetto da attività agricole e di sviluppo. Lo scavo del recinto settentrionale fornirà senza dubbio altri indizi sulla storia di Hambledon Hill nel Neolitico e sulla storia del suo abbandono.

La chimica dell'aglio e della cipolla

Agli insoliti composti solforati che sono responsabili dell'odore dell'aglio e che fanno lacrimare chi affetta le cipolle si devono anche le notevoli proprietà terapeutiche da lungo tempo attribuite a queste due piante

di Eric Block

Il mondo è stato sempre diviso in due: da una parte coloro che amano l'aglio e la cipolla, dall'altra coloro che li detestano. Nel primo gruppo potremmo mettere i faraoni egizi, che furono sepolti assieme a piccole sculture di argilla e legno, raffiguranti bulbi d'aglio e di cipolla, cosicché i pasti consumati nell'aldilà fossero saporiti. Potremmo annoverarvi anche gli ebrei che vagabondarono per 40 anni nel deserto del Sinai, ricordando con nostalgia «i pesci che mangiavamo in abbondanza quando eravamo in Egitto, e le zucche e i meloni, e i porri, le cipolle e l'aglio.» Potremmo includervi Sydney Smith, un saggista del XIX secolo, la cui *Ricetta per l'insalata* contiene questo consiglio: «Lasciate che rimangano nascosti nella pignatta pochi atomi di cipolla, appena percepiti: ravviveranno tutto il sapore».

Nel secondo gruppo, avverso all'aglio e alla cipolla, dovremmo annoverare i sacerdoti egizi che, secondo Plutarco, «si astenevano dal mangiare cipolla che... non è adatta né per il digiuno né per le celebrazioni, perché nel primo caso provoca sete, nel secondo lacrime in coloro che vi partecipano». Dalla stessa parte si situerebbero anche i greci antichi, che consideravano volgare l'odore dell'aglio e della cipolla e proibivano l'entrata nel tempio di Cibele a coloro che ne avevano mangiato. Spregiatore di aglio e cipolla è anche Bottom, un personaggio del *Sogno di una notte di mezza estate,* che istruisce la sua compagnia d'attori a «non mangiare né cipolla né aglio, perché dobbiamo avere un alito gradevole».

Nel gruppo degli estimatori di aglio e cipolla si possono includere, per motivi professionali, anche i chimici: infatti sono sempre stati attratti da sostanze con odori forti, sapori piccanti ed effetti fisiologici marcati. Le loro ricerche, protrattesi per oltre un secolo, hanno stabilito che, nel momento in cui si taglia un bulbo di cipolla o d'aglio, si libera un certo numero di molecole organiche di peso molecolare basso, contenenti atomi di zolfo con legami raramente osservabili in natura. Queste molecole sono molto reattive: si modificano spontaneamente in altri composti organici solforati, che prendono parte a ulteriori trasformazioni. Esse mostrano, inoltre, una notevole gamma di effetti biologici, di cui la proprietà di produrre lacrimazione è solo un esempio. Alcuni estratti di aglio e cipolla sono antibatterici e antimicotici. Altri sono antitrombotici, cioè impediscono alle piastrine del sangue di formare trombi, ossia aggregati di piastrine e di molecole di fibrina (una sostanza proteica): in altri termini, impediscono al sangue di coagularsi.

L'aglio e la cipolla sono rappresentanti della famiglia delle gigliacee: i loro nomi botanici sono rispettivamente *Allium sativum* e *Allium cepa (allium* forse deriva dalla parola celtica *all,* che significa pungente). Entrambe le specie sono tra le più antiche piante coltivate: la loro origine, molto probabilmente nell'Asia centrale, risale alla preistoria e per millenni sono state utilizzate nella medicina popolare. Il Codice Ebers, un papiro egizio di medicina, datato al 1550 a.C. circa, fornisce più di 800 formule terapeutiche, di cui 22 menzionano l'aglio come rimedio efficace nei riguardi di numerosi disturbi, tra cui affezioni cardiache, dolori di testa, morsicature, infestazioni di vermi e tumori.

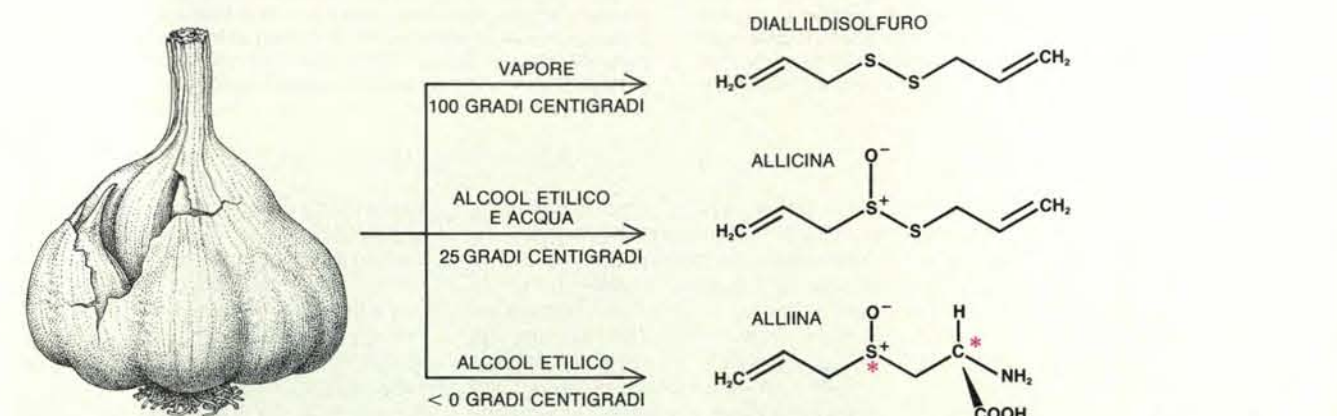
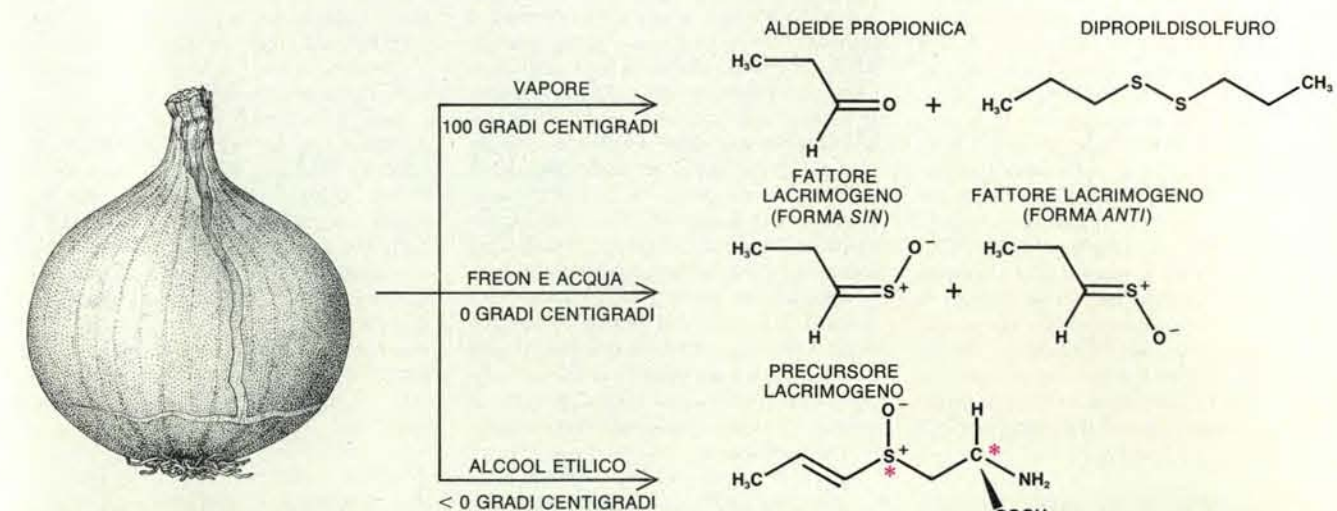
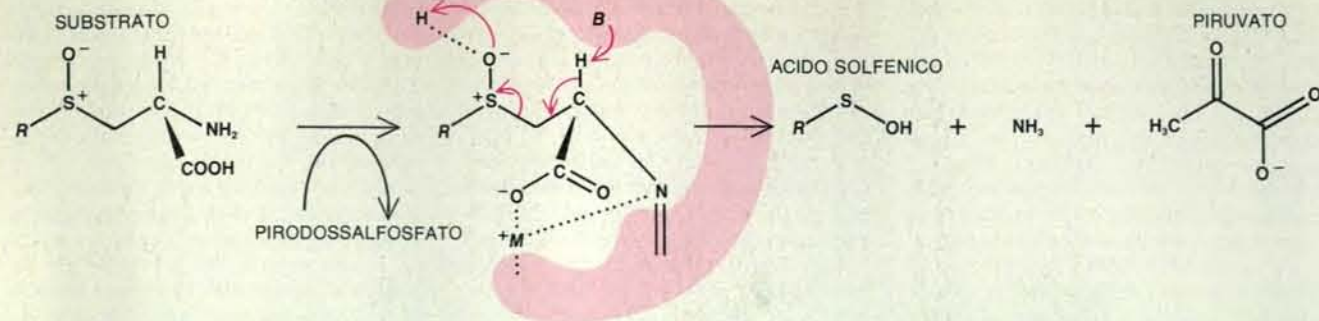
Gli egizi non furono i soli ad apprezzare aglio e cipolla. Ippocrate e Aristofane raccomandavano l'aglio per le sue proprietà medicamentose. Plinio il Vecchio citava numerose utilizzazioni terapeutiche sia per l'uno sia per l'altra. Dioscoride, medico presso l'esercito romano nel I secolo d.C., prescriveva l'aglio come vermifugo. Durante i primi giochi olimpici in Grecia, sembra che l'aglio venisse consumato dagli atleti come stimolante.

In India l'aglio è stato usato come lozione antisettica per lavare ferite e ulcere. In Cina il tè di cipolle è stato a lungo raccomandato per la febbre, il mal di testa, il colera e la dissenteria. La medicina popolare è spesso intrecciata con la leggenda, come nel caso dell'«aceto dei quattro ladri». Si racconta che, nel 1721, quattro criminali fossero stati reclutati per seppellire i morti durante una terribile pestilenza a Marsiglia. I quattro becchini si rivelarono immuni dalla malattia; loro segreto era una bevanda, costituita da aglio macerato nel vino, che divenne immediatamente famosa come *vinaigre des quatre voleurs* e ancora oggi è reperibile in Francia.

Assieme a queste prescrizioni popolari, è emerso più di recente anche un attestato scientifico. Secondo una serie di ricerche, aglio e cipolla hanno mostrato di possedere una blanda azione antibiotica. Nel 1858 Pasteur scoprì le proprietà antibatteriche dell'aglio. Più di recente, si dice che Albert Schweitzer, in Africa, abbia fatto uso di aglio per il trattamento della dissenteria amebica. Nelle due guerre mondiali l'aglio fu usato come antisettico nella prevenzione della cancrena. In ricerche di laboratorio si può evidenziare che il suo succo, diluito fino a una parte su 125 000, inibisce la crescita dei batteri dei generi *Staphylococcus, Streptococcus, Vibrio* (compreso *V. cholerae*) e *Bacillus* (compresi *B. typhosus, B. dysenteriae* e *B. enteritidis*). Inoltre, mostra un largo spettro di attività contro i funghi zoopatogeni e contro molti ceppi di lievito, compresi alcuni che provocano la vaginite.

Secondo un altro filone di prove aglio e cipolla spiccano per la loro efficacia contro le trombosi. Anche in questo caso le prove sono vecchie e nuove. In Francia, un tempo, ai cavalli affetti da trombosi alle zampe si somministravano aglio e cipolla. Più di recente, e cioè nel 1979, G. S. Sainani e collaboratori del B. J. Medical College dell'Università di Poona, in India, hanno pubblicato i risultati di uno studio epidemiologico compiuto su tre popolazioni che consumavano quantitativi differenti d'aglio e cipolla. I soggetti erano vegetariani della comunità giainista, che mangiavano aglio e cipolla in quantità elevata (almeno 50 grammi di aglio e 600 grammi di cipolle alla settimana), oppure in piccole quantità (non più di 10 grammi d'aglio e 200 grammi di cipolle alla settimana), o mai per tutta la vita.

Il gruppo di coloro che si astenevano dal consumare aglio e cipolla presentò tempi brevi nella coagulazione del sangue. Inoltre, presentò il più elevato livello plasmatico di fibrinogeno. (Parte del processo di coagulazione del sangue è dovuta alla trasformazione del fibrinogeno in fibrina.) Già negli studi compiuti negli anni settanta era stato notato che gli oli estratti dall'aglio e dalla cipolla inibivano l'aggregazione delle piastrine. Dunque, le credenze popolari attorno alle due piante sembravano acquistare credito.

In che modo aglio e cipolla producono i loro effetti? La risposta deve essere ricercata a livello molecolare, tra le sostanze contenute in essi. Una delle ricerche chimiche di più vecchia data fu compiuta nel 1844 dal chimico tedesco Theodor Wertheim ed ebbe come soggetto l'aglio. Wertheim attribuiva l'interesse nei riguardi di questa pianta «principalmente alla presenza di un corpo liquido, contenente zolfo: l'olio d'aglio. Tutto ciò che si sa di questa sostanza si limita ad alcune semplici constatazioni sul prodotto puro, che si ottiene per distillazione in corrente di vapore dai bulbi di *Allium sativum.* Poiché finora i le-

Il tipo di composti solforati che si possono estrarre dall'aglio dipende dalle condizioni d'estrazione. La tecnica più brutale è la distillazione in corrente di vapore, vale a dire la bollitura dell'aglio seguita dall'estrazione dei composti dal vapore condensato: questo metodo fornisce il diallildisolfuro *(in alto).* Una tecnica più raffinata consiste nell'utilizzare come solvente l'alcool etilico a temperatura ambiente: si ottiene così l'ossido del diallildisolfuro, l'allicina *(al centro),* che è la causa del tipico odore dell'aglio. Con una tecnica ancor più delicata, che utilizza l'alcool etilico puro a una temperatura al di sotto dello zero, si ottiene l'alliina, molecola dotata di isomeria ottica, ossia con forme chimiche aventi strutture speculari rispetto agli atomi di zolfo e di carbonio *(asterischi).* (Sono possibili quattro forme, ma solo una si trova realmente nell'aglio.) Grazie a un enzima, l'alliina si può trasformare in allicina. Di ogni molecola viene indicato solo lo scheletro carbonioso.

Anche i composti solforati estratti dalla cipolla dipendono dalle condizioni d'estrazione. La distillazione in corrente di vapore fornisce l'aldeide propionica e il dipropildisolfuro *(in alto).* Grazie al Freon, un solvente mescolato con acqua a zero gradi centigradi, si ottiene il «fattore lacrimogeno» *(al centro),* la sostanza che fa lacrimare chi affetta una cipolla. Questo fattore si presenta in due forme isomere, designate con i prefissi *sin* e *anti*: la forma *sin* è prevalente. Utilizzando alcool etilico come solvente, a temperature al di sotto dello zero, si ottiene infine il «precursore lacrimogeno» *(in basso),* isomero strutturale dell'alliina; in altre parole, l'alliina e questo precursore differiscono solamente per la formula di struttura. Nella cipolla un enzima trasforma il precursore lacrimogeno nel fattore lacrimogeno.

L'enzima allinasi catalizza la trasformazione, nell'aglio e nella cipolla, di parecchi composti solforati. Cosa del massimo rilievo, questo enzima, nell'aglio, agisce sull'alliina, mentre nella cipolla agisce sul precursore lacrimogeno. Nello schema è raffigurato un caso di catalisi su un substrato generico. (Per esempio, se *R* è un gruppo allile, C_3H_5, il substrato è l'alliina.) Sul substrato agisce un cofattore, il piridossalfosfato, che gli fa formare un complesso con l'enzima; il legame comprende l'interazione elettrostatica del substrato e di uno ione metallico (M^+). Un gruppo basico *(B)*, presente sull'enzima, sposta poi un protone, o idrogenione, dal substrato, provocandone la demolizione e liberando un acido solfenico, RSOH, con ammoniaca e piruvato. Una reazione chimica è in sostanza una trasformazione di legami chimici all'interno delle molecole, che può essere simboleggiata dal movimento di coppie di elettroni: i movimenti più probabili sono indicati da frecce *(in colore)*.

gami di zolfo sono stati studiati poco, le ricerche su questa sostanza promettono di fornire risultati utili alla scienza».

Wertheim utilizzava la tecnica della distillazione in corrente di vapore. Metteva l'aglio in acqua bollente e il vapore che si liberava dal recipiente conteneva piccoli quantitativi di olio d'aglio, la cui distillazione forniva alcune sostanze volatili di odore assai forte. Egli propose perciò il nome di «allile» (da *Allium*) per il radicale dell'idrocarburo contenuto nell'olio e quello di *schwefelallyl* (in italiano «solfoallile») per i composti volatili. Il termine «allile» si usa ancora oggi: si riferisce a gruppi aventi formula di struttura $CH_2{=}CHCH_2$, o formula bruta C_3H_5. Numerosi composti comprendenti un allile hanno un odore pungente.

Nel 1892 un altro ricercatore tedesco, il chimico F. W. Semmler, applicò la distillazione in corrente di vapore agli spicchi d'aglio, producendo uno o due grammi di un olio dal pessimo odore per ogni chilogrammo di prodotto di partenza. A sua volta l'olio fornì diallildisolfuro ($C_6H_{10}S_2$ o, più precisamente, $CH_2{=}CHCH_2SSCH_2CH{=}CH_2$), accompagnato da minori quantità di dialliltri- e dialliltetrasolfuro *(si veda l'illustrazione in alto a pagina 75)*. Sempre mediante distillazione in corrente di vapore, da cinque tonnellate di cipolle si otteneva un olio abbastanza diverso, che conteneva l'aldeide propionica (C_2H_5CHO) assieme a numerosi composti solforati, come il dipropildisolfuro ($C_6H_{12}S_2$).

La successiva fondamentale scoperta nella chimica dell'aglio e della cipolla fu fatta nel 1944 da Chester J. Cavallito e collaboratori alla Sterling-Winthrop Chemical Company di Rensselaer, nello stato di New York. Questi ricercatori stabilirono che con metodi meno brutali della distillazione in corrente di vapore si ottenevano sostanze abbastanza diverse. Cavallito trattò con alcool etilico quattro chilogrammi d'aglio a temperatura ambiente e, alla fine, ottenne sei grammi di un olio la cui formula era $C_6H_{10}S_2O$ e che aveva proprietà antibatteriche e antimicotiche. Era più potente della penicillina e della sulfaguanidina nei riguardi di *Bacillus typhosus*; negli altri casi si rivelava meno efficace della penicillina.

L'olio di Cavallito è, dal punto di vista chimico, ossido di diallildisolfuro, la principale sostanza che Semmler, mezzo secolo prima, aveva isolato per distillazione in corrente di vapore. La sua formula è $CH_2{=}CHCH_2S(O)SCH_2CH{=}CH_2$, il che rende abbastanza difficile il suo nome: allil-2-propentiosolfinato. Va detto che la nomenclatura chimica è piuttosto complicata, ma assai precisa nella caratterizzazione di una molecola. Ogni parte di un termine chimico riporta la struttura di una sezione dello scheletro carbonioso della molecola oppure segnala l'interruzione dello scheletro carbonioso da parte di atomi diversi, come gli atomi di zolfo. Nell'allil-2-propentiosolfinato il 2 indica che il doppio legame (=) tra carbonio e carbonio parte dal secondo atomo di carbonio, numerato a partire dal punto di attacco dello zolfo. Le parentesi che racchiudono un atomo o un gruppo di atomi indicano che quell'atomo o quel gruppo di atomi non fa parte della catena principale della molecola.

Cavallito denominò però questa nuova sostanza con un nome più semplice: allicina. Si tratta di un liquido incolore, chimicamente instabile, che giustifica appieno l'odore dell'aglio, molto più di quanto facciano i diallilsolfuri. Esso è oggetto negli Stati Uniti di ben due brevetti registrati sotto il nome di Cavallito, ma il suo impiego in campo clinico come agente antibatterico è stato abbandonato dopo un breve periodo di sperimentazione a causa dell'odore.

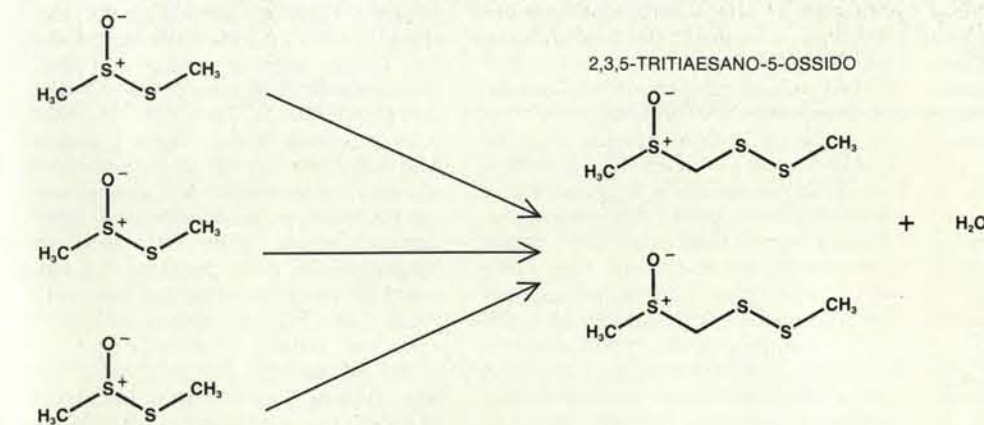
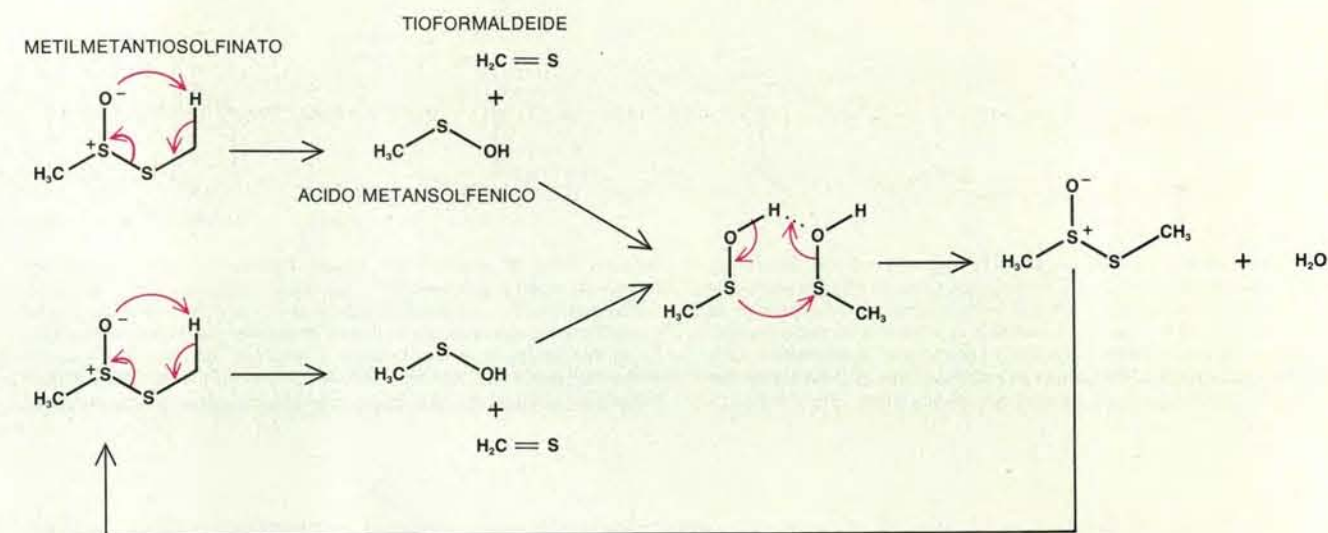
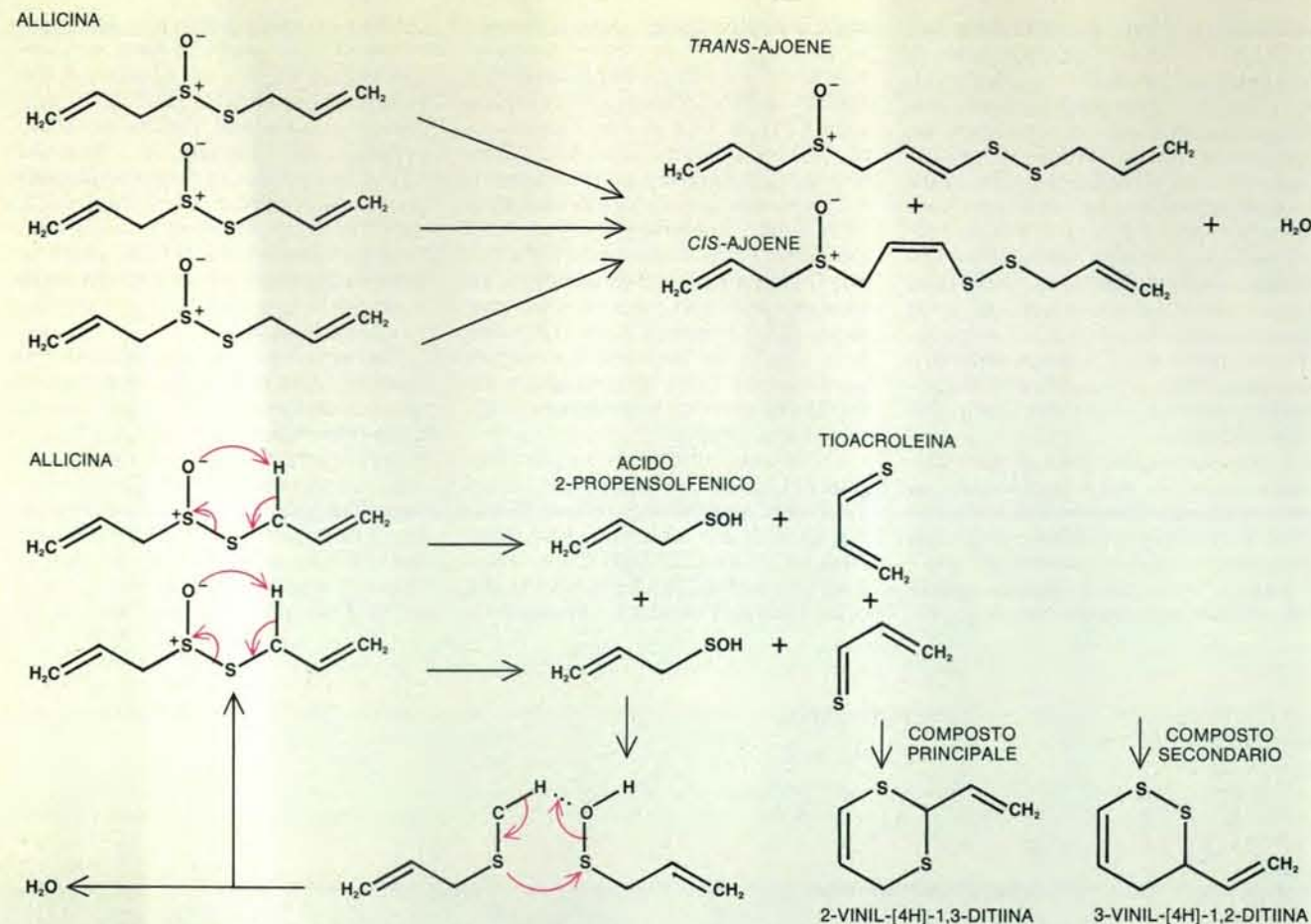
L'allicina è sì responsabile dell'odore dell'aglio, ma un bulbo d'aglio non emana praticamente alcun odore finché non viene tagliato o schiacciato. Nel 1948, Arthur Stoll ed Ewald Seebeck della Sandoz a Basilea trovarono la spiegazione di questo fenomeno: l'allicina si sviluppa nell'aglio quando un enzima la forma a partire da un precursore inodoro, che Stoll e Seebeck hanno identificato come (+)-S-allil-L-cisteinsolfossido, con formula $CH_2{=}CHCH_2S(O){-}CH_2CH(NH_2)COOH$. (Il segno + e la lettera L indicano una particolare disposizione spaziale per l'atomo di zolfo e per quello di carbonio attaccato all'azoto.) Evidentemente il taglio o lo schiacciamento dell'aglio permette all'enzima, chiamato allinasi, di venire in contatto con il precursore dell'allicina.

Stoll e Seebeck diedero il nome di alliina a questo precursore che forma circa lo 0,24 per cento del peso di un tipico bulbo d'aglio. L'alliina può formarsi allorché un allile e un atomo di ossigeno si attaccano all'atomo di zolfo nell'amminoacido cisteina. Essa può anche essere estratta dall'aglio, ma questa operazione deve svolgersi in condizioni chimiche assai blande. Per successiva cristallizzazione si ottengono cristalli aghiformi finissimi, incolori e inodori.

L'alliina è una molecola dalle peculiari proprietà: in particolare, è stata la prima sostanza naturale a presentare isomeria ottica, dovuta a forme speculari rispetto all'atomo di zolfo e a quello di carbonio. L'isomeria ottica si ha quando una molecola si presenta in forme speculari e la natura favorisce una delle due forme. Una soluzione della sostanza si dimostra perciò in grado di far ruotare un fascio di luce polarizzata. Nella alliina, sono possibili configurazioni speculari a livello sia dello zolfo sia del carbonio. Sotto l'influenza dell'allinasi, l'alliina si decompone in acido 2-propensolfenico *(si veda l'illustrazione nella pagina a fronte)*. L'enzima agisce preferenzialmente sull'isomero dell'alliina designato con (+), ossia sulla forma che provoca la rotazione di un fascio di luce polarizzata in senso orario. A sua volta l'acido 2-propensolfenico dimerizza, ossia si accoppia con una seconda molecola dello stesso acido per dare l'allicina.

Mentre le ricerche sulla chimica dell'aglio erano in pieno svolgimento, si avviarono anche quelle sulla cipolla. Nel 1961 il biochimico finlandese Artturi Virtanen (che aveva ricevuto nel 1945 il premio Nobel per la chimica, grazie alle ricerche sull'allevamento degli animali) dimostrò che le cipolle contengono *trans*-(+)-S-(1-propenil)-L-cisteinsolfossido, un isomero di posizione dell'alliina *(si veda l'illustrazione in basso a pagina 75)*. In altre parole, il contenuto chimico è identico a quello della alliina: solo la struttura differisce. (Precisamente differisce nella posizione di un doppio legame che, come indica l'1 nel nome del composto, parte direttamente dallo zolfo.) Il *trans*-(+)-S-(1-propenil)-L-cisteinsolfossido è il «precursore lacrimogeno» (PL); l'enzima allinasi, presente nella cipolla, lo trasforma nel «fattore lacrimogeno» (FL), ossia in quella sostanza che provoca lacrimazione in chi affetta una cipolla.

La formula bruta del fattore lacrimogeno è C_3H_6SO, che corrisponde a più di 50 diverse formule di struttura. Virtanen ipotizzò che la formula di struttura fosse $CH_3CH{=}CHS(O)H$ e non la formula alternativa $CH_3CH{=}CHS{-}O{-}H$, in cui l'atomo di ossigeno è situato nella catena principale carboniosa della molecola. (Questi composti sono entrambi denominati acido *trans*-1-propensolfenico.) Nel frattempo W. F. Wilkins, laureando alla Cornell University, aveva proposto la formula di struttura $C_2H_5CH{=}SO$, che corrisponde al nome propantial-S-ossido. Dieci anni dopo, nel 1971, M. H. Brodnitz e J. V. Pascale della International Flavors and Fragrances Company a Union Beach, nello stato del New Jersey, hanno confermato questa ipotesi.

Nell'aglio, quindi, una allinasi trasforma l'alliina in allicina, il componente odoroso dell'aglio. Nella cipolla una allinasi trasforma il precursore lacrimogeno nel fattore lacrimogeno. Queste non sono le sole reazioni controllate dai due enzimi che, sia nell'aglio sia nella cipolla, agiscono su numerose molecole: parlando in gergo chimico, si dice che questi enzimi hanno un certo numero di substrati. Tutti questi substrati sono costi-

La decomposizione del metilmetantiosolfinato, omologo o versione semplificata dell'allicina, avviene secondo passaggi che hanno permesso di chiarire le vie metaboliche che portano all'allicina stessa. Lungo una via *(in alto)* il metilmetantiosolfinato si decompone in acido metansolfenico e in tioformaldeide; poi due molecole di acido metansolfenico si combinano per rigenerare una molecola di metilmetantiosolfinato. Lungo un'altra via *(in basso)*, tre molecole di metilmetantiosolfinato si condensano per produrre 2,3,5-tritiaesano-5-ossido.

La decomposizione dell'allicina procede lungo diverse vie. In una *(in alto)*, tre molecole di allicina si combinano, producendo due molecole di una sostanza chiamata ajoene. Il meccanismo è stato suggerito da studi compiuti sul metilmetantiosolfinato. L'ajoene è un antitrombotico, efficace almeno quanto l'aspirina nel prevenire l'aggregazione delle piastrine e perciò la coagulazione del sangue. Esso si presenta in due forme, designate *trans* e *cis;* la *cis* è lievemente meno efficace. Lungo un'altra serie di reazioni *(in basso)* l'allicina si autodecompone, formando acido 2-propensolfenico e tioacroleina, entrambe sostanze molto reattive. Per autocondensazione di due molecole di acido 2-propensolfenico si rigenera una molecola di allicina; per autocondensazione di due molecole di tioacroleina si formano due tipi di composto ciclico mediante un processo chimico, la reazione di Diels-Alder. I due nuovi composti ciclici hanno una blanda azione antitrombotica.

tuiti da sostanze contenenti zolfo, sintetizzate nell'aglio e nella cipolla mediante sequenze chimiche che partono dall'amminoacido solforato cisteina. Da essi, le allinasi formano diversi acidi solfenici, RSOH, dove R indica un radicale: o l'allile ($CH_2{=}CHCH_2$), o l'1-propenile ($CH_3CH{=}CH$), o il metile (CH_3) o il propile (C_3H_7). I sottoprodotti delle reazioni sono il piruvato ($CH_3C(O)COO^-$) e l'ammoniaca (NH_3).

Secondo le più recenti ricerche, le reazioni richiedono la partecipazione di una sostanza addizionale o cofattore; il piridossalfosfato. Evidentemente cofattore e substrato interagiscono e pertanto il substrato viene trasformato in una forma attivata. Un gruppo basico presente nell'enzima (ossia un gruppo che può catturare protoni) inizia a questo punto la liberazione di acido solfenico. Da parte loro gli acidi solfenici sono estremamente instabili e vanno incontro spontaneamente a ulteriori trasformazioni.

Le mie ricerche sulla chimica dell'aglio e della cipolla hanno avuto inizio nel 1971 con una più approfondita esplorazione delle proprietà dell'allicina. I miei collaboratori e io, all'Università del Missouri a Saint Louis, abbiamo cominciato a studiare la trasformazione chimica del metilmetantiosolfinato, $CH_3S(O)SCH_3$. Il composto è l'omologo più semplice dell'allicina: da un lato presenta un gruppo chimico (S(O)H) fondamentale per la chimica dell'allicina; dall'altro, lo scheletro carbonioso è più semplice di quello dell'allicina. Assieme a John O'Connor ho scoperto due processi chimici insoliti *(si veda l'illustrazione a pagina 77)*. In uno di essi, per decomposizione del metilmetantiosolfinato, si ottiene acido metansolfenico, CH_3SOH, e tioformaldeide, $CH_2{=}S$. A loro volta, due molecole di acido metansolfenico, sostanza notevolmente reattiva, si combinano (con la perdita di una molecola d'acqua) e si riforma così una molecola di metilmetantiosolfinato. Nel secondo processo da noi studiato, invece, il metilmetantiosolfinato subisce una reazione di autocondensazione, producendo il 2,3,5-tritiaesano-5-ossido: $CH_3S(O)CH_2SSCH_3$.

A distanza di dodici anni, la nostra ricerca si è rivelata importante nel chiarire la struttura e il modo di formazione del fattore antitrombotico dell'aglio. Mahendra K. Jain e Roger W. Crecely dell'Università del Delaware, in collaborazione con Rafael Apitz-Castro e Maria R. Cruz dell'Istituto venezuelano di ricerche scientifiche di Caracas, hanno prodotto parecchi estratti d'aglio, particolarmente attivi nel prevenire l'aggregazione delle piastrine del sangue. L'estratto più attivo aveva formula bruta $C_9H_{14}S_3O$. In stretta collaborazione con i nostri colleghi delle università del Delaware e del Venezuela, Saleem Ahmad e io, alla State University of New York ad Albany, siamo riusciti a chiarire la struttura del composto, la quale è $CH_2{=}CHCH_2S(O)CH_2CH{=}CHSSCH_2CH{=}CH_2$, ossia 4,5,9-tritiadodeca-1,6,11-trien-9-ossido. La denominazione da noi data al composto è «ajoene», dalla parola spagnola *ajo*, che significa aglio.

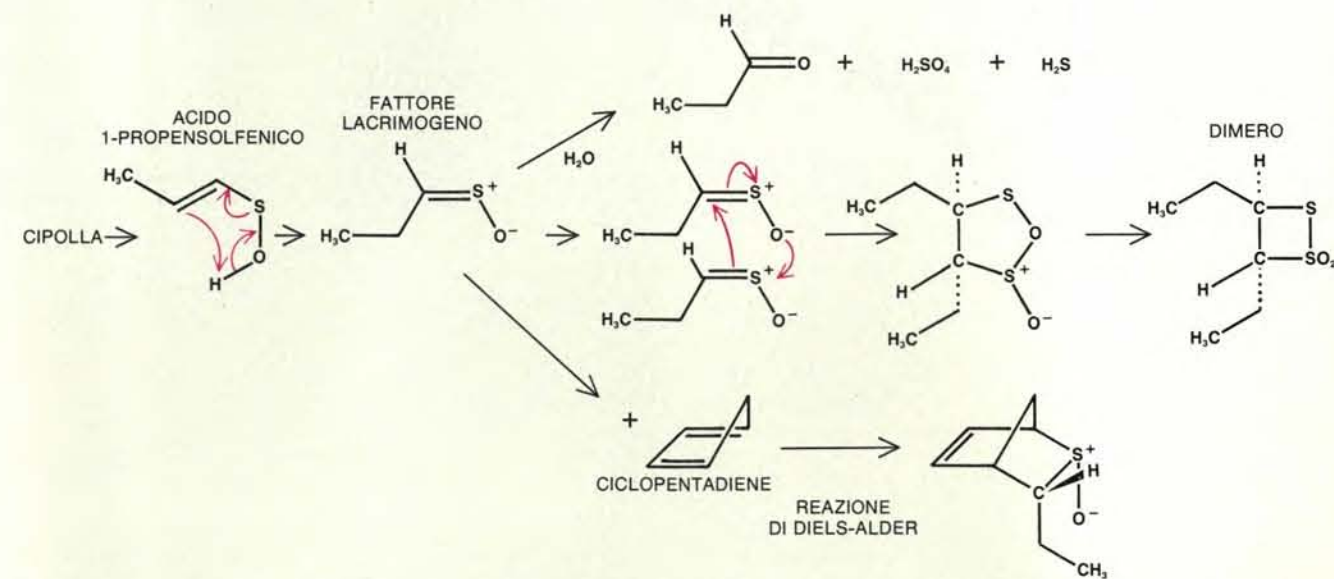
Le mie prime ricerche, compiute sull'autocondensazione del metilmetantiosolfinato, hanno suggerito che l'ajoene potesse formarsi per autocondensazione dell'allicina. Abbiamo verificato questa ipotesi scaldando semplicemente l'allicina con un miscuglio d'acqua e di un solvente organico come l'acetone *(si veda l'illustrazione nella pagina a fronte)*. Negli esperimenti condotti in seguito si è potuto dimostrare che l'ajoene come fattore antitrombotico è potente almeno quanto l'aspirina. Gli studi compiuti da gruppi di ricerca delle università del Delaware e del Venezuela, in collaborazione con James Catalfamo del New York State Department of Health ad Albany, fanno pensare, infine, che l'ajoene agisca inibendo i recettori del fibrinogeno sulle piastrine. Più precisamente, vi può essere un'interazione dei gruppi idrocarburici solfossigenati e disolfurici dell'ajoene con gruppi complementari sotto l'aspetto chimico e presenti sulla superficie delle piastrine, i quali potrebbero altrimenti legarsi al fibrinogeno. Nuovi esperimenti oggi in corso dovrebbero stabilire l'eventuale utilità farmacologica dell'ajoene.

Un secondo aspetto della chimica del metilmetantiosolfinato, omologo dell'allicina, si è anche rivelato interessante. Ho fatto rilevare poc'anzi che la decomposizione del metilmetantiosolfinato produce tioformaldeide ($CH_2{=}S$). Sembra che anche per l'allicina abbia luogo il medesimo tipo di processo. In particolare, la decomposizione dell'allicina produce tioacroleina, $CH_2{=}CH{-}CH{=}S$, un composto fortemente reattivo, di colore blu zaffiro. Hans Bock, dell'Università di Francoforte, ha dimostrato che la tioacroleina dimerizza, formando due composti ciclici, che abbiamo trovato nell'aglio, nel rapporto in cui Bock li avrebbe previsti. La dimerizzazione procede mediante una reazione di Diels-Alder, in cui un'unità tetraatomica di una molecola si combina con un'unità biatomica di un'altra molecola per formare un anello esaatomico. Le reazioni di Diels-Alder sono tra le più importanti in chimica organica.

Per quanto concerne la cipolla, rimaneva un problema da risolvere: il fattore lacrimogeno era stato identificato come $C_2H_5CH{=}SO$, ossia propantial-S-ossido, ma questa molecola ha due isomeri. Nel tentativo di scoprire quale dei due è contenuto nella cipolla, Larry Revelle, Robert E. Penn e Ali Bazzi hanno studiato il problema nel mio laboratorio. Per estrarre il fattore lacrimogeno hanno sminuzzato alcune cipolle congelate, hanno utilizzato un solvente (Freon), hanno prodotto un residuo eliminando il solvente a -78 gradi centigradi e hanno distillato il residuo sotto vuoto a -20 gradi centigradi. Mediante due tecniche spettroscopiche molecolari indipendenti (la spettroscopia a microonde e la spettroscopia a risonanza magnetica nucleare) sono poi riusciti a stabilirne la struttura. Si tratta in gran parte di *sin*-propantial-S-ossido il cui isomero *anti* è presente solo in traccia. Nella forma *sin* il gruppo etilico (C_2H_5), a un'estremità della catena carboniosa della molecola, è prossimo all'atomo di ossigeno presente all'altra estremità della catena.

Un ulteriore problema è stato chiarito nel mio laboratorio da due esperimenti. Nell'aglio le allinasi producono acidi solfenici; nella cipolla producono propantial-S-ossidi, distinti dagli acidi solfenici. In particolare, i propantial-S-ossidi appartengono alla classe di composti detti solfine. Se le solfine si formassero direttamente, dovrebbero avvenire processi chimici completamente diversi: questa conclusione è strana perché un enzima, di solito, catalizza un solo tipo di reazione, non parecchi.

Nel primo esperimento, Penn ha stabilito mediante spettroscopia che la struttura dell'acido metansolfenico (il più semplice tra gli acidi solfenici) è $CH_3S{-}O{-}H$ e non $CH_3S(O)H$. Poi, in un secondo esperimento, Penn e io abbiamo trovato che quando l'acido *trans*-1-propensolfenico (il fattore lacrimogeno di Virtanen) viene preparato con metodi chimici ristruttura rapidamente la propria molecola trasformandosi in *sin*-propantial-S-ossido. Se si ammette (basandosi sul primo esperimento) che l'acido *trans*-1-propensolfenico abbia struttura $CH_3CH{=}CHS{-}O{-}H$, anziché $CH_3CH{=}CHS(O)H$, la ristrutturazione della molecola può essere dedotta facilmente come un trasporto interno di idrogeno *(si veda l'illustrazione in questa pagina)*. Si può concludere pertanto che la fase iniziale nella formazione del fattore lacrimogeno della cipolla produce effettivamente un acido solfenico (acido 1-propensolfenico), che in seguito si trasforma rapidamente nel vero fattore lacrimogeno: il *sin*-propantial-S-ossido.

Il fattore lacrimogeno è anch'esso fortemente reattivo. In laboratorio può essere idrolizzato, fornendo (insieme ad altre sostanze) acido solforico. Può di-

Le vie che portano al fattore lacrimogeno sono complesse. Nell'immediato precursore del fattore, l'acido 1-propensolfenico, il gruppo SOH partecipa a un doppio legame. La vicinanza favorisce un trasporto interno di idrogeno *(frecce in colore)* e, quindi, la formazione del fattore. Questo può subire un'idrolisi *(in alto)*, formando aldeide propionica, acido solforico e acido solfidrico; può combinarsi con se stesso, cioè dimerizzarsi *(al centro)*, formando un curioso anello a quattro atomi; infine, può essere bloccato in una struttura biciclica *(in basso)*.

merizzare, formando un dimero nella cui strana struttura è inserito un anello a quattro atomi. Come Alan Wall e io abbiamo trovato, questo dimero può dar luogo a una reazione di Diels-Alder con il ciclopentadiene, molecola estremamente reattiva, contenente una unità a quattro atomi, il «diene». La reazione blocca la struttura *sin* del fattore lacrimogeno in una intelaiatura molecolare rigida, fatta di due anelli uniti tra loro.

Le proprietà chimiche del fattore lacrimogeno chiariscono l'efficacia dei metodi usati in cucina per attenuare il disagio di chi deve affettare cipolle. La sua volatilità viene ridotta fortemente tenendo in frigorifero la cipolla. Inoltre, sbucciando la cipolla sotto l'acqua corrente, esso viene facilmente asportato, in quanto è idrosolubile.

Perché la natura ha incorporato nell'aglio e nella cipolla questo apparato chimico per fabbricare l'allicina e il fattore lacrimogeno? Poiché l'allicina è antimicotica oltre che antibiotica, potrebbe proteggere la pianta dell'aglio dall'eventuale decomposizione del bulbo, provocata da funghi. E poiché il fattore lacrimogeno della cipolla è irritante e ripugnante per taluni animali, potrebbe essere anch'esso importante per la sopravvivenza della pianta.

Resta insoluta la questione riguardante la proprietà antitrombotica di alcune molecole presenti nell'aglio. I miei collaboratori e io non siamo mai riusciti a scoprire ajoene o composti ciclici antitrombotici nella polvere disidratata d'aglio; né l'abbiamo trovata nelle pillole, negli olii, negli estratti o in altri preparati brevettati a base d'aglio. La probabile spiegazione è che la fabbricazione della maggior parte di tali prodotti comincia con la distillazione dell'aglio in corrente di vapore. Per ora, gli effetti benefici attribuiti all'aglio si possono ottenere nel modo migliore facendo uso di aglio fresco. Naturalmente, il trattamento autonomo che alcune persone praticano su di sé con preparati a base di aglio o di cipolla non deve sostituire un'accurata diagnosi e terapia medica. Basta un minimo di buon senso per rendersene conto; nel caso questo non bastasse si può ricorrere all'olfatto. L'ingestione di aglio e cipolla lascia un ricordo duraturo, perché i composti a base di zolfo, introdotti nel flusso sanguigno, trovano una via d'uscita nell'aria espirata e nella traspirazione. Virtù e vizi dell'aglio sono riassunti in modo mirabile da Sir John Harrington nell'opera *The Englishman's Doctor*, scritta nel 1609:

L'aglio ha poi la proprietà
di salvare dalla morte;
sopportalo, anche se rende
l'alito disgustoso,
e non disprezzarlo come taluni,
convinti che faccia soltanto
bruciare gli occhi, bere smodatamente
e maleodorare.

Perché saltano le balene

Sembra che i grandi balzi fuori dall'acqua di questi e altri cetacei abbiano scopi ben precisi: essi sarebbero correlati con aspetti sociali della vita di questi animali, in particolare con la comunicazione

di Hal Whitehead

Quando i cetacei balzano fuori dall'acqua, quasi certamente assistiamo a uno dei più straordinari movimenti che un animale possa compiere. Questo «aprirsi un varco» nella superficie del mare - come lo definirono i balenieri del XVIII e del XIX secolo - è ancora oggi analizzato dai ricercatori che si interessano a questo fenomeno. Se si considerano la mole e il peso notevoli che un cetaceo deve sollevare, ci si può chiedere quali siano le ragioni di un simile comportamento.

I balzi fuori dall'acqua *(breach)* forniscono l'unica occasione di poter vedere un cetaceo per intero e hanno sempre ispirato una grande varietà di commenti. Così, infatti, nel 1839 J. N. Reynolds raccontava per i lettori del «Knickerbocker» le avventure dei balenieri nel Pacifico: «Di tanto in tanto un enorme corpo informe balza fuori dal suo elemento, ricadendo con pesante impatto; tutta la scena è una ridicola caricatura di agilità così come lo sarebbe quella di grassi notabili di paese che si cimentassero in una danza scozzese.» Per Herman Melville, l'autore di *Moby Dick*, si trattava di una manifestazione sublime: «emergendo alla massima velocità dagli abissi più profondi - scriveva - il capodoglio lancia tutta la propria massa corporea nel puro elemento dell'aria, sollevando una montagna di schiuma scintillante e mostrando dove si trova a distanza di sette miglia o anche più. In quegli istanti, le onde infrante e spostate dalla sua mole formano sul suo corpo una specie di criniera».

I primi balenieri, che andavano alla caccia delle loro prede con lente imbarcazioni a vela, avevano molte occasioni per osservare le balene che cercavano di catturare. Per anni e anni, gli aneddoti raccontati da questi uomini fornirono molti particolari utili sui balzi fuori dall'acqua e su altri comportamenti tipici delle balene e degli altri cetacei. Le loro spiegazioni, che peccavano un poco di antropocentrismo, li interpretavano come movimenti legati all'alimentazione, alla necessità di stirarsi, al gioco, al bisogno di sfuggire all'inseguimento dei pesci spada o, semplicemente, come «atto di sfida» presumibilmente rivolto agli stessi balenieri.

Nel corso degli ultimissimi anni, le osservazioni scientifiche condotte su cetacei in mare aperto hanno cominciato a fornire utili dati quantitativi su molti aspetti del loro comportamento, compresi i balzi fuori dall'acqua. Roger Payne, del World Wildlife Fund statunitense, e i suoi collaboratori, dopo lunghi studi sulla balena australe *(Eubalaena australis)* al largo della penisola argentina di Valdés, hanno fornito molte spiegazioni. Altri importanti studi sono stati condotti sulla balena grigia della California *(Eschrichtius robustus)*, al largo della Baja California, da Kenneth S. Norris dell'Università della California a Santa Cruz e una serie di osservazioni, come quelle di James D. Darling, dell'Università della California a Santa Cruz, di Peter Tyack, della Woods Hole Oceanographic Institution, e di altri, è stata realizzata sulle megattere *(Megaptera novaeangliae)* al largo di Hawaii. Il mio lavoro ha riguardato principalmente le megattere che frequentano nei mesi estivi l'Atlantico nordoccidentale al largo di Terranova e durante i mesi invernali si concentrano, invece, nella zona di Silver Bank nelle Indie Occidentali.

Per capire i balzi fuori dall'acqua dei cetacei sono indispensabili osservazioni a lungo termine, poiché il fenomeno non è generalizzato ed è raro vedere degli esemplari che si abbandonano a questa manifestazione. Di conseguenza, sono necessari molti anni di studio per poter raccogliere un numero appena sufficiente di dati. Sotto questo aspetto, la ricerca svolta a Silver Bank è stata particolarmente importante. Le megattere provenienti dall'Atlantico nordoccidentale si raccolgono in quella zona nei mesi invernali per accoppiarsi e partorire. La densità della loro popolazione è di circa un esemplare per chilometro quadrato. Molte balzano fuori dall'acqua: durante le nostre traversate, che si estendevano per circa 200 chilometri da una parte all'altra del Silver Bank e che compivamo per valutare bene l'entità della popolazione, abbiamo osservato balzi fuori dall'acqua in circa il 20 per cento dei gruppi da noi rilevati (e che erano abitualmente costituiti da una a quattro megattere).

Una megattera che balza fuori dall'acqua solleva una biomassa pari a quella di 485 persone del peso medio di 68 chilogrammi ciascuna. Le megattere più grosse raggiungono lunghezze attorno ai 15 metri e un peso di 33 tonnellate. I loro salti e quelli di altri cetacei che li praticano variano da un'emersione completa fuori dall'acqua a un'emersione lenta, in cui fuoriesce solo una metà del corpo. In oltre un quarto dei balzi effettuati, almeno il 70 per cento del corpo dell'animale si proietta nell'aria, mentre più raramente si può vedere l'animale completo al di sopra del pelo dell'acqua. Rispetto alla superficie del mare, le megattere balzano con tutte le angolazioni possibili fino a un massimo di 70 gradi.

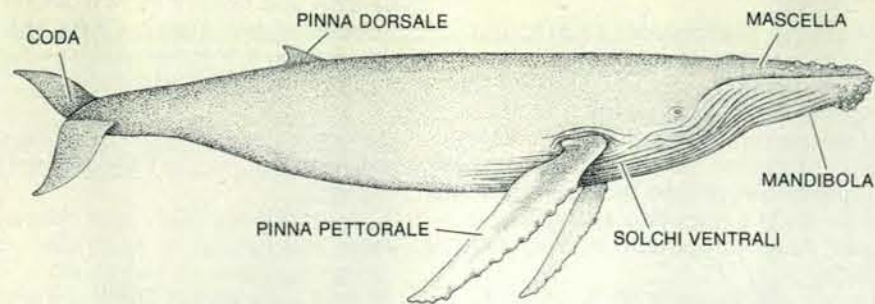
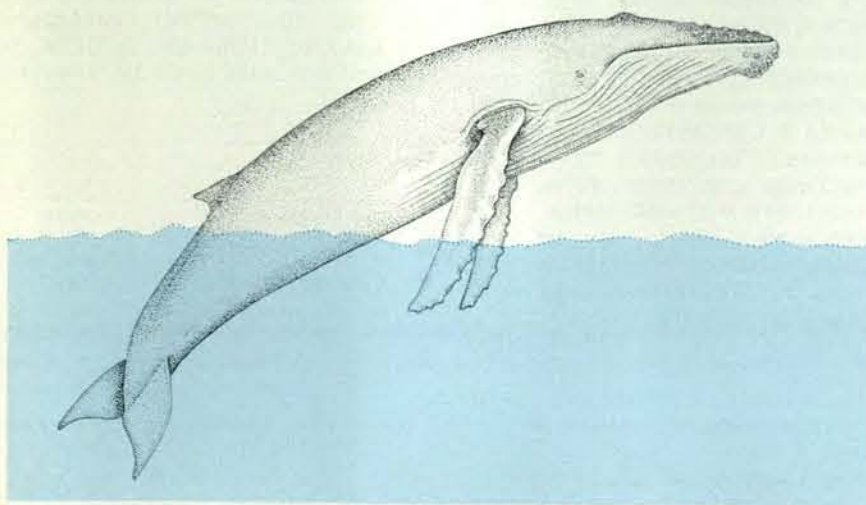
Payne ha osservato i balzi delle balene australi dall'alto degli scogli oppure sorvolando il mare con piccoli aeroplani. La balena nuota orizzontalmente fino ad acquistare una sufficiente velocità, poi alza la testa e solleva la coda. Questi movimenti trasformano la sua quantità di moto da orizzontale in verticale e l'animale può così uscire dall'acqua. In virtù della progressione orizzontale esso può effettuare un balzo in aria anche in acque profonde pochi metri.

I cetacei eseguono altri movimenti che assomigliano superficialmente al balzo fuori dall'acqua. In una di queste manovre (che gli anglosassoni chiamano *lunge*) l'animale sporge dall'acqua con non più del 40 per cento del corpo; il movimento può essere eseguito orizzontalmente, verticalmente o con varie angolazioni fra questi due estremi; il corpo può essere con il dorso o con il ventre rivolti verso l'alto, oppure girato su un fianco. Spesso, in queste occasioni, si osserva l'animale che chiude le mascelle e inghiotte il plancton o piccoli pesci. Pertanto, questo tipo di movimento è generalmente associato all'alimentazione. Tuttavia è anche possibile osservare delle megattere che lo effettuano per superarsi l'un l'altra all'interno di grossi gruppi, per esempio quando da due a dieci maschi entrano in concorrenza per accaparrarsi una femmina del branco. Sembra allora che l'emersione dell'animale dall'acqua sia involontaria e conseguente a manovre compiute sott'acqua. Per contro il balzo in aria sembra avere sempre uno scopo ben determinato.

Un altro movimento intenzionale dei cetacei al di sopra della superficie del mare è quello del delfino, in cui l'animale compie una serie di balzi in avanti suborizzontali *(porpoising)* mentre procede rapidamente. Robert W. Blake, dell'Università della Columbia Britannica, ha calcolato che, con questi balzi, un delfino o comunque un cetaceo di piccole dimensioni riduce al minimo la resistenza dell'acqua. Egli ha anche dimostrato che i grossi cetacei non avrebbero alcuna utilità a prodursi in questo genere di movimento e, in effetti, io non ho mai osservato delle megattere esibirsi in esso.

I balzi fuori dall'acqua rientrano in due categorie. Nei balzi con avvitamento *(true breach)*, il cetaceo esce dall'acqua su un fianco, effettua con il corpo una torsione, agitando le pinne pettorali, e ricade sul dorso. Nell'altra categoria di balzi, invece, l'animale rimane con il dorso rivolto verso l'alto per tutto il tempo in cui emerge dall'acqua e ricade con una spanciata *(belly flop)*. Le megattere eseguono balzi del primo tipo nell'80 per cento dei casi e, per l'altro 20 per cento, balzi con spanciate. È in questo tipo di balzo che è più facile osservare una megattera emettere dallo sfiatatoio un pennacchio di vapore che condensa. Secondo Payne il balzo con ricaduta sul ventre sarebbe per il cetaceo altrettanto doloroso che le spanciate che gli esseri umani prendono tuffandosi male; però è anche vero che lo sfiatatoio viene lasciato fuori dall'acqua per un tempo maggiore rispetto a quanto accade nel balzo con avvitamento. Potrebbe quindi costituire l'alternativa preferita quando l'animale desidera respirare durante un balzo.

I balzi, alle volte, vengono eseguiti in sequenza. Un cetaceo può compierne a intervalli di 40 secondi circa per qualche minuto. Fra le megattere dell'Atlantico nordoccidentale la lunghezza della sequenza era in media di 9,4 balzi (comprendendo nella media anche i casi di balzi singoli). In genere, tutti i balzi sembravano effettuati da un unico soggetto, come si è verificato a Silver Bank quando abbiamo contato una sequenza di 130 balzi in 75 minuti, probabilmente eseguita dallo stesso individuo. All'interno di una sequenza, i balzi sono tendenzialmente tutti dello stesso tipo: o un balzo con spanciata dopo l'altro, o un balzo con avvitamento e ricaduta sul dorso dopo l'altro. Sia fra le megattere sia fra le balene australi, un soggetto che

Questa balenottera ripresa nel Pacifico, presso Hawaii, è una megattera *(Megaptera novaeangliae)*, che esegue un balzo con avvitamento, emergendo dall'acqua su un fianco, effettuando una torsione del corpo e ricadendo sul dorso. Nell'altro tipo di balzo, molto più raro, il cetaceo rimane sempre con il dorso verso l'alto e ricade sul ventre con una spanciata. Circa l'80 per cento dei balzi osservati è del primo tipo.

La megattera è stata oggetto della maggior parte delle osservazioni che l'autore ha compiuto sui balzi fuori dall'acqua dei cetacei. I soggetti da lui studiati insieme con i suoi collaboratori trascorrono l'estate al largo di Terranova e l'inverno nella zona di Silver Bank, nelle Indie occidentali. Le interazioni sociali delle megattere sono probabilmente più importanti nel corso dell'inverno, che rappresenta la stagione durante la quale si svolgono gli accoppiamenti e i parti.

Ogni balzo inizia con l'emersione delle megattere dall'acqua, con tutte le possibili angolazioni fino a 70 gradi rispetto alla superficie del mare. Nel balzo in alto, che è quello con avvitamento, l'animale ricade sul dorso; in quello in basso, ricade con una spanciata. Spesso, in questo caso viene emesso un pennacchio di vapore che condensa e dà l'impressione che la megattera scelga questo balzo (in verità raro) quando vuole respirare; l'esercizio consente, infatti, di mantenere lo sfiatatoio fuori dall'acqua per un tempo maggiore rispetto a quanto avviene nel balzo con avvitamento.

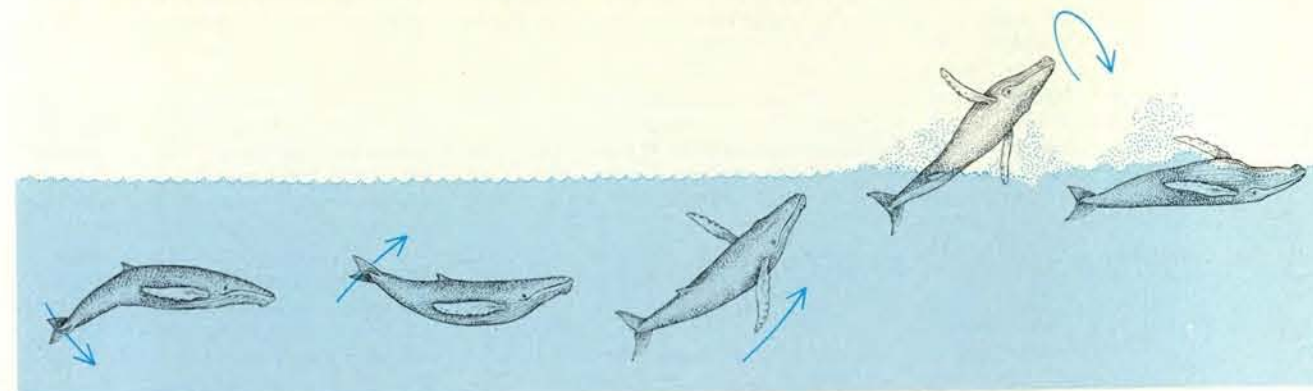
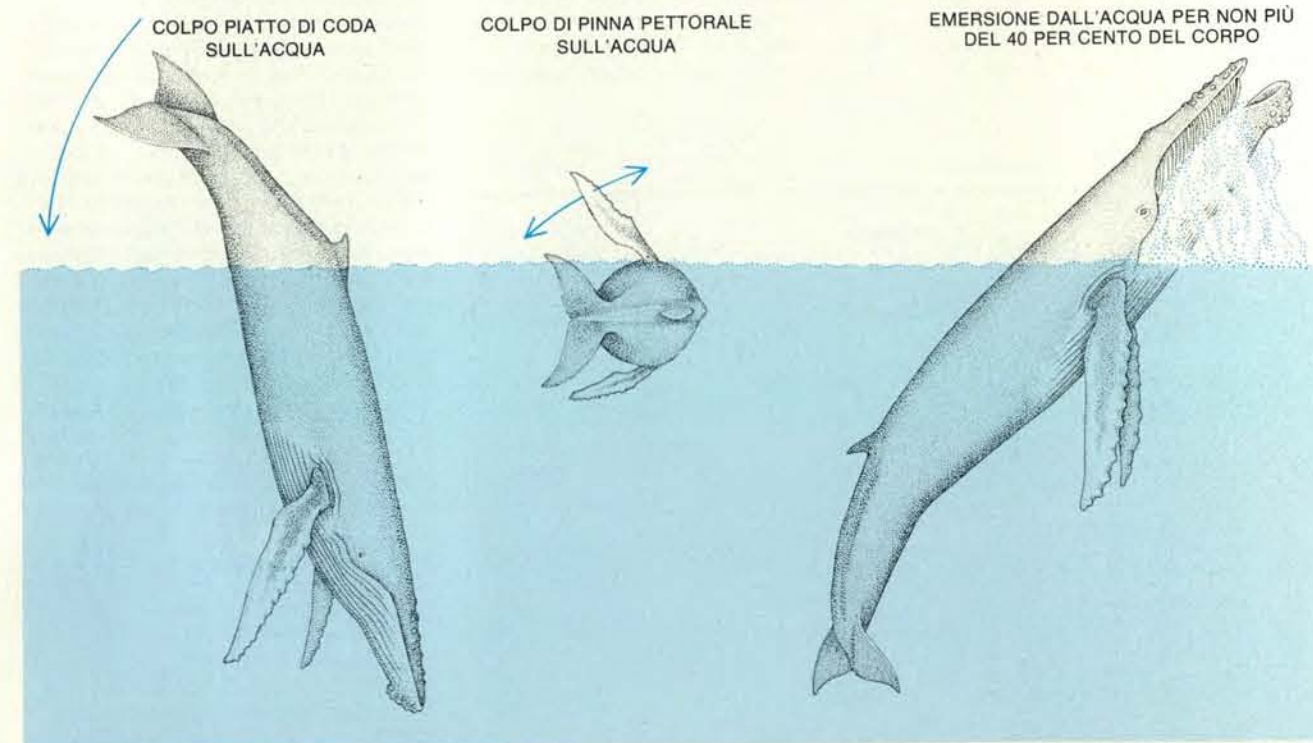
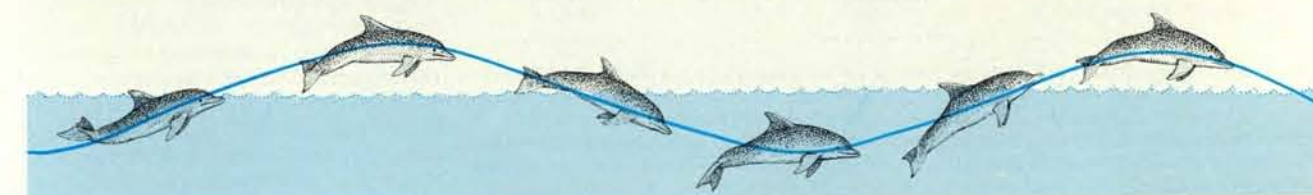
compia dei balzi successivi tende a fare emergere una parte sempre più ridotta del proprio corpo con il procedere della sequenza. Come è facilmente intuibile, l'animale dà l'impressione di stancarsi.

Quanta energia consuma un cetaceo quando esegue un balzo in aria e quanta potenza sviluppa quando si stacca dalla superficie del mare? Utilizzando delle misurazioni che ho ricavato da fotogrammi che ritraggono dei cetacei che compiono balzi in aria, sono riuscito a simulare il processo in un piccolo calcolatore. In un balzo in cui la maggior parte del corpo emerge dall'acqua con una angolazione di circa 35 gradi, una megattera adulta erompe dalla superficie del mare a una velocità di circa 28 chilometri all'ora. Essendo questa la massima velocità raggiungibile da quell'animale, un balzo completamente fuori dall'acqua rappresenta il massimo sfruttamento che una megattera fa della propria potenza propulsiva.

L'energia necessaria per effettuare un balzo di questo tipo ammonta approssimativamente a 2500 chilocalorie. Il tasso metabolico di una megattera allo stato di riposo è di circa 300 000 chilocalorie al giorno. Quindi l'energia spesa nel corso di un singolo balzo corrisponde a poco meno di un centesimo del fabbisogno calorico minimo giornaliero dell'animale: energia che si traduce in poco più di due chilogrammi e mezzo di pesce cappellano, un componente importante della dieta delle megattere, che spesso ingeriscono pesci di questa specie a porzioni di un quintale alla volta. Pertanto un balzo fuori dall'acqua non costituisce un evento particolarmente significativo nel bilancio energetico giornaliero delle megattere. Una sequenza di 20 balzi o più consuma però una consistente quantità di energia e non deve perciò sorprendere che i balzi perdano progressivamente in potenza.

È meno facile spiegare perché un cetaceo si produca nei balzi fuori dall'acqua. Lo studio del comportamento dei grossi cetacei è stato paragonato all'astronomia. Infatti, l'osservatore ha una visione fuggevole del soggetto, spesso a grande distanza; non può fare esperimenti; infine, deve continuamente trarre conclusioni da una messe di dati insolitamente inadeguati. In queste condizioni, uno dei modi di indagare sul ruolo di una attività è di esaminare il contesto in cui si svolge.

Ho trascorso centinaia di ore a bordo di piccole barche a vela seguendo gruppi di megattere intente alle loro attività quotidiane. Questo lavoro, unitamente alle osservazioni di Payne e altri, sta fornendo un quadro abbastanza chiaro delle circostanze in cui questi cetacei si producono nei balzi, anche se non intende stabilire un insieme di regole fisse riguardo al fenomeno. Nello studio del comportamento degli animali progrediti, di norma non è possibile acquisire queste certezze. La miglior cosa che si possa fare è indicare alcune tendenze che appaiono statisticamente significative; queste fanno comprendere che il movimento a balzi è un comportamento prevalentemente associato a un'interazione sociale, forse nella comunicazione o nel gioco (nel caso degli esemplari più giovani).

Spesso le megattere eseguono dei balzi quando un gruppo con due o più esemplari si scinde in due gruppi distinti o quando due di questi gruppi (a volte costituiti da singole megattere) si riuniscono in uno solo. Un balzo viene notato spesso entro 15 minuti da quando l'ani-

L'esecuzione di un balzo inizia quando la megattera, nuotando più o meno parallelamente alla superficie, acquista velocità. Essa solleva la coda e volge la testa in alto, modificando la quantità di moto da orizzontale a verticale. Si tratta di un balzo con avvitamento.

Fra i movimenti che si collegano ai balzi fuori dall'acqua vi sono la percussione della superficie dell'acqua con un colpo netto da parte della coda e il sollevamento di una pinna pettorale che viene poi sbattuta sull'acqua. Sembra che si tratti sempre di movimenti con un preciso intento. Un movimento chiaramente non intenzionale, che risulta da manovre compiute sott'acqua, è quello che fa emergere solo una parte del corpo *(lunge)*. Spesso, dopo una quindicina di minuti in cui si assiste ai due movimenti citati, l'animale esegue un balzo.

Un delfino compie una serie di balzi in avanti suborizzontali mentre nuota a velocità elevata. Per un cetaceo di piccola mole, come questo tursiope, tale tipo di movimento riduce al minimo la resistenza dell'acqua. Al contrario, esso non avrebbe particolare efficacia nei cetacei di grossa mole.

male ha cominciato a percuotere in piatto la superficie del mare mediante la coda *(lobtailing)*; il fenomeno è spesso associabile a quel movimento delle pinne pettorali *(flippering)* che imita il battito delle ali degli uccelli, con la pinna che viene sollevata e poi sbattuta sull'acqua, e ad altre manifestazioni. Christopher W. Clark, della Rockefeller University, e Payne hanno osservato comportamenti analoghi fra le balene australi.

È un fatto indicativo e apparentemente contraddittorio che le megattere eseguano un minor numero di balzi durante l'estate, anche se in questa stagione, più spesso che in inverno, i gruppi si scindono e si riformano. Ma, nella stagione invernale, le megattere si accoppiano e partoriscono e queste interazioni sociali sono probabilmente più importanti di quelle abituali dell'estate. Per questa ragione, la frequenza dei balzi è correlata non soltanto con il numero di interazioni sociali che si realizzano, ma anche con l'importanza che queste hanno nella vita dell'animale.

Se si raffrontano le frequenze dei balzi nelle diverse specie di cetacei, si rileva una ulteriore correlazione di questo tipo di movimento con l'attività sociale. Nell'analizzare questo aspetto, ho compilato una tabella in cui il rapporto fra massa e cubo della lunghezza del corpo esprime la «rotondità» (si tratta, in realtà, più che di un corpo rotondo, di un corpo meno slanciato). Si dovrebbe pensare che le specie dal corpo meno slanciato siano meno portate ai balzi, essendo sfavorite sotto l'aspetto idrodinamico; invece le osservazioni dimostrano che esse li praticano con maggiore frequenza.

Le balene australi, le balene grige e le megattere, che sono le tre specie con corpo meno slanciato meglio studiate, convergono durante la stagione invernale nelle zone di riproduzione tradizionali, in cui però ben raramente si nutrono, limitandosi a sfruttare l'energia contenuta nei loro spessi strati adiposi. In queste aree, le interazioni sociali sono frequenti, talvolta molto intense, ed è proprio in esse che i balzi vengono eseguiti in numero elevato.

Per contro, non sembra che la balenottera azzurra *(Balaenoptera musculus)*, la balenottera comune *(B. physalus)* e la balenottera boreale *(B. borealis)* - tutte di corporatura più slanciata - frequentino zone di riproduzione determinate e, durante i mesi invernali, rimangono disperse. Questa strategia riduce probabilmente il loro dispendio energetico netto e quindi non richiede spessi strati adiposi. Per formare le coppie ricorrono verosimilmente a richiami sonori a bassa frequenza oppure a un sistema sociale monogamo. Comunque hanno probabilmente scarse interazioni sociali strette.

I sistemi sociali della balena della Groenlandia *(Balaena mysticetus)*, della balenottera di Bryde *(Balaenoptera edeni)* o della balenottera minore *(B. acutorostrata)* non sono molto noti, ma l'impressione generale di chi ha potuto osservare da vicino questi cetacei è che, fra loro, le specie più socializzanti eseguono dei balzi con una maggior frequenza. Il capodoglio *(Physeter catodon)*, un cetaceo odontoceto che si esibisce frequentemente nei balzi in aria, ha un sistema sociale particolarmente complesso.

SPECIE		RAPPORTO MASSA/CUBO DELLA LUNGHEZZA	FREQUENZA DEI BALZI
MEGATTERA *(MEGAPTERA NOVAEANGLIAE)*	15 METRI	10,6	MOLTO ELEVATA
BALENA AUSTRALE *(EUBALAENA AUSTRALIS)*		16,2	ELEVATA
BALENA GRIGIA DELLA CALIFORNIA *(ESCHRICHTIUS ROBUSTUS)*		14,3	ELEVATA
CAPODOGLIO (MASCHIO/FEMMINA) *(PHYSETER CATODON)*		10,7/19,1	ELEVATA
BALENA DELLA GROENLANDIA *(BALAENA MYSTICETUS)*		26,7	BASSA
BALENOTTERA DI BRYDE *(BALAENOPTERA EDENI)*		6,1	BASSA
BALENOTTERA MINORE *(BALAENOPTERA ACUTOROSTRATA)*		12,3	MOLTO BASSA
BALENOTTERA COMUNE *(BALAENOPTERA PHYSALUS)*		4,0	ESTREMAMENTE BASSA
BALENOTTERA AZZURRA *(BALAENOPTERA MUSCULUS)*		6,3	PRATICAMENTE NULLA
BALENOTTERA BOREALE *(BALAENOPTERA BOREALIS)*		3,6	PRATICAMENTE NULLA

I balzi in aria e la «rotondità» del corpo (in realtà la forma tozza del corpo) sembrano collegati: più un cetaceo ha forma poco slanciata più è probabile che si esibisca in questi esercizi. La rotondità è data dal rapporto fra la massa del corpo e il cubo della sua lunghezza. In linea di massima, i soggetti più slanciati si producono con minor frequenza nei balzi, anche se l'idrodinamica di questi ultimi sembrerebbe favorirli. La correlazione nasce dal fatto che le specie dalla forma meno slanciata si impegnano spesso nel tipo di attività sociale collegata ai balzi, particolarmente quando si riuniscono nelle zone di riproduzione durante i mesi invernali. In questo periodo non si nutrono molto, sfruttando per lo più le loro riserve di grasso. I cetacei più slanciati sono, invece, molto meno sociali e probabilmente si nutrono senza interruzione per tutto l'anno.

Quali altri indizi emergono dalle ricerche sul contesto in cui i cetacei compiono i loro balzi? Un risultato inatteso, ottenuto da una serie di studi indipendenti, è che la frequenza dei balzi è direttamente proporzionale alla velocità del vento. Non si tratta di un brusco aumento durante le burrasche, nel qual caso si potrebbe pensare che il cetaceo emerga per respirare senza emettere un pennacchio di vapore dagli sfiatatoi, ma piuttosto di un aumento graduale, coincidente con velocità del vento del tutto moderate. Payne ha ipotizzato che i cetacei ricorrano ai balzi per una comunicazione sonora (il tonfo del rientro) quando il rumore del vento e delle onde attutisce i suoni che essi normalmente emettono.

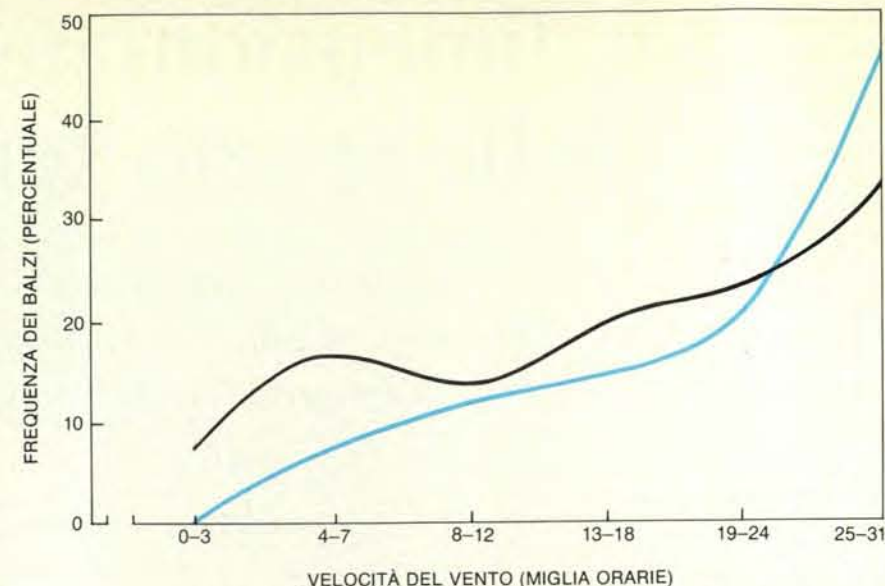
Payne ha fatto un'altra scoperta che lo ha indotto a pensare che i balzi abbiano una funzione di segnalazione. Egli ha notato che, fra le balene australi, essi sono un fenomeno che si autopropaga. In altri termini, le probabilità che una balena cominci a compiere dei balzi in aria aumentano quando anche le balene vicine li stanno eseguendo. Incuriosito da questi dati, ho eseguito una rudimentale analisi spettrale in alcune delle nostre traversate a Silver Bank.

I risultati ottenuti suggeriscono che le megattere che compiono balzi formino gruppi estesi con un diametro di circa 10 chilometri; una megattera aveva più probabilità di effettuare balzi quando si trovava a una distanza inferiore ai 10 chilometri da altre megattere intente allo stesso esercizio. Al culmine della stagione i gruppi potevano comprendere anche 100 individui, di cui 10 o 15 individui soltanto si esibivano nei balzi. In condizioni favorevoli essi sarebbero stati in grado di captare il rumore di un balzo a pochi chilometri di distanza. Questi risultati sembravano corroborare l'ipotesi, avanzata da Payne, di una funzione di segnalazione del balzo. Se le altre megattere vedono o odono il rumore di un balzo, ciò significa che l'informazione è stata trasmessa. Il messaggio è - se non altro - che un'altra megattera del gruppo ha eseguito un balzo.

Ma i balzi in aria sono un modo efficace per trasmettere altre forme di messaggi? Per gli osservatori in superficie essi costituiscono un evento spettacolare e assai fragoroso, ma quando una megattera balza fuori dall'acqua la maggior parte delle altre è sommersa. Anche in acque limpidissime, il limite di visibilità è di circa 50 metri, mentre in condizioni favorevoli il suono può propagarsi nell'acqua di mare a distanze ben più notevoli. L'interrogativo che ci si pone è se, con questi balzi fuori dall'acqua, una megattera possa produrre suoni più forti, almeno entro alcune fasce di frequenza, di quanto lo consentano i suoi mezzi vocali. Non esistono informazioni sufficienti sulla intensità dei suoni che un balzo in aria produce sott'acqua, e non si sa neppure se le megattere, durante questi balzi, cerchino di rendere massime le loro emissioni sonore.

Il balzo fuori dall'acqua potrebbe anche rappresentare una manifestazione di aggressività, una specie di sfida, o di dimostrazione di forza, o di rituale nel corteggiamento. L'ipotesi dell'aggressività non è molto fondata, in quanto i cetacei sono animali dalla pelle liscia, con molti strati protettivi, e quindi è difficile vedere come un balzo potrebbe produrre danni a un immaginario nemico, a meno che questo non sia molto più piccolo dell'aggressore. Mi risulta che una volta una megattera, durante una serie di balzi in aria al largo di Terranova, sia caduta sopra un'imbarcazione, ma è presumibile che si sia trattato di un incidente più che di una manifestazione di aggressività da parte di quell'animale. Sono stato per mesi in alto mare, su piccole imbarcazioni, ma non ho mai avuto l'impressione che le migliaia di balzi che ho osservato fossero diretti contro di noi. Per di più, per dimostrare la propria aggressività, la megattera può ricorrere molto più efficacemente a un forte colpo di coda.

Un cetaceo che esegua un balzo completamente fuori dall'acqua fa sfoggio della sua massima potenza di fronte a qualsiasi altro animale congenere, a portata di vista o di udito. Di conseguenza, questa esibizione potrebbe essere utile ai fini del corteggiamento, come sfida o come dimostrazione di forza. Una femmina potrebbe scegliersi il partner sessuale almeno in parte in base alla potenza con cui questo balza fuori dall'acqua o alla capacità che esso ha di prolungare la manifestazione di forza o sonora per tutta una sequenza di balzi. Un maschio dimostrerebbe così la propria forza e il proprio vigore e quindi (indirettamente) la propria idoneità genetica.

Analoghe correlazioni renderebbero i balzi in aria utili come sfida o come dimostrazione di forza diretta contro altri maschi concorrenti nel corteggiamento di una determinata femmina.

È stata riscontrata una correlazione positiva fra i balzi compiuti dai cetacei fuori dall'acqua e la velocità del vento. Essa viene presentata nel grafico in base ai dati raccolti dall'autore nel 1978 *(in nero)* e nel 1980 *(in colore)*. Roger Payne, del WWF statunitense, crede che i balzi fuori dall'acqua si intensifichino con l'aumento della velocità del vento, poiché contribuiscono a migliorare la comunicazione quando il rumore delle onde attenua le normali emissioni vocali.

Quando maschi di balene australi o di megattere sono impegnati in questi episodi di competizione, si notano frequentemente dei balzi in aria.

Si deve anche prendere in considerazione il concetto, per quanto un poco confuso, del gioco. Quando vediamo un animale eseguire un movimento che non sia traducibile immediatamente in una funzione palese, tendiamo ad affermare che si stia trattando di un gioco. Di conseguenza, il concetto è diventato una categoria in cui vengono compresi tutti i comportamenti che non si è in grado di spiegare altrimenti: i balzi fuori dall'acqua hanno spesso subìto questa sorte. Di recente il gioco è stato studiato attentamente da alcuni biologi e studiosi del comportamento animale e oggi viene generalmente considerato come una valida (anche se difficilmente definibile) categoria comportamentale. Se il compiere balzi fuori dall'acqua è un'attività importante per i cetacei e se il modo in cui questi balzi vengono eseguiti incide sulla loro efficacia, vi sono buone ragioni d'ordine selettivo per le quali i giovani, o forse anche gli adulti, debbano praticare questo «gioco».

I balzi in aria presentano gran parte delle caratteristiche di altre attività che gli studiosi del comportamento animale definiscono gioco: sono abituali nei contesti sociali, sono spesso praticati dai giovani e, in molti casi, non suggeriscono una funzione palese. Alcuni ricercatori hanno ipotizzato che, in altri giovani animali, lo scopo del gioco sia quello di contribuire allo sviluppo della muscolatura; nei giovani cetacei, i balzi in aria potrebbero adempiere a questa funzione.

I balzi più spettacolari sono quelli effettuati dagli esemplari più giovani. I piccoli delle balene australi, delle balene grige o delle megattere cominciano a saltare fuori dall'acqua poche settimane dopo la nascita, con balzi vigorosi, talvolta ripetuti a lungo. A Silver Bank, i giovani si esibivano in questo esercizio più frequentemente degli adulti. In effetti sarebbe piuttosto eccezionale che degli adulti si dedicassero regolarmente a questa energica attività per gioco. Sembra perciò improbabile che il gioco sia, fra le balene adulte, la principale motivazione dei balzi fuori dall'acqua.

I risultati che ho riportato e le ipotesi che ho discusso non indicano, per i balzi fuori dall'acqua, un'unica funzione ben determinata. Le prove ricavate suggeriscono che questa attività abbia in realtà parecchie funzioni. Sebbene vi siano forti correlazioni con la socialità e i balzi abbiano caratteristiche distintive come segnali di vigore fisico, non è stato possibile verificare concretamente né un'ipotesi né l'altra.

La mia valutazione soggettiva è che i balzi in aria servano spesso ad accentuare altre comunicazioni visive o acustiche, quasi come un punto esclamativo di natura fisica. I cetacei praticano i balzi in aria proprio come noi alziamo la voce, gesticoliamo o ci agitiamo per dar maggior risalto a un messaggio. Ma come chi origlia, anche gli osservatori umani si lasciano di solito sfuggire il significato del messaggio e notano unicamente i suoi aspetti più appariscenti.

Le dimensioni nascoste dello spazio-tempo

Lo spazio-tempo, comunemente considerato tetradimensionale, può avere anche sette dimensioni in più, e sono proprio le strutture a undici dimensioni che potrebbero portare all'unificazione delle forze della natura

di Daniel Z. Freedman e Peter van Nieuwenhuizen

Il 29 maggio 1919 l'ombra di un'eclisse totale di Sole si estendeva attraverso l'Atlantico dall'Africa occidentale al Brasile settentrionale. Spedizioni organizzate dal governo britannico su suggerimento di Sir Arthur Stanley Eddington si apprestavano a osservare le stelle in prossimità del disco oscurato del Sole. Uno degli obiettivi principali di Eddington era la verifica di una nuova teoria della gravità, formulata da Einstein quattro anni prima, meglio conosciuta con il nome di relatività generale. In questa teoria Einstein avanzava la sorprendente pretesa intellettuale che la geometria dell'universo fosse determinata dalla materia e dall'energia in esso contenute. Più esattamente, secondo la relatività generale, lo spazio e il tempo sono intimamente connessi in una struttura matematica tetradimensionale chiamata spazio-tempo, mentre la forza di gravità viene spiegata come un effetto della cosiddetta curvatura intrinseca dello spazio-tempo.

Gli osservatori dell'eclisse si accingevano a verificare direttamente uno degli effetti previsti nel contesto dello spazio-tempo curvo di Einstein. Secondo la relatività generale, il cammino percorso dalla luce emessa da stelle in prossimità del Sole verrebbe curvato dall'attrazione gravitazionale solare, cosicché quando il disco solare si avvicina a una stella, questa dovrebbe apparire spostata dalla sua solita posizione celeste. Per verificare la teoria fu necessario attendere un'eclisse solare perché solo in questa occasione si sarebbero potute vedere le stelle vicine al Sole. Le osservazioni dell'eclisse resero Einstein famoso in tutto il mondo in quanto le stelle risultarono spostate esattamente delle entità previste e venne così confermato senza ombra di dubbio il successo del modo einsteiniano di affrontare geometricamente la gravità.

Anche se la relatività generale ha a che fare solo con la geometria a quattro dimensioni, le geniali ricerche di Einstein aprirono la strada ad applicazioni sempre più audaci della sua idea fondamentale. Nello stesso anno in cui il concetto di un universo tetradimensionale veniva confermato dalle osservazioni astronomiche, Theodor Franz Eduard Kaluza, un giovane studioso e libero docente, praticamente sconosciuto, dell'Università di Königsberg (l'odierna città di Kaliningrad, in Unione Sovietica), inviò a Einstein un saggio in cui proponeva di aggiungere alle quattro dimensioni dello spazio-tempo una quinta dimensione spaziale.

Kaluza introduceva una quinta dimensione per poter dare una spiegazione unificata di tutte le forze conosciute della natura. A quel tempo si conoscevano solo due forze fondamentali: la gravitazione, descritta dalla relatività generale, e l'elettromagnetismo, descritto dalla teoria di James Clerk Maxwell e altri. Le due forze sembrano profondamente differenti; per esempio, tutte le particelle sono soggette alla gravità, ma solo le particelle cariche sono soggette all'elettromagnetismo. Nel 1914 Gunnar Nordström dell'Università di Helsingfors (l'odierna Helsinki) aveva tentato di dare una descrizione unificata delle due forze apparentemente diverse dimostrando che entrambe nascono da una forma pentadimensionale dell'elettromagnetismo, ma il suo metodo venne abbandonato perché non riusciva a spiegare la curvatura della luce nei pressi del Sole. Kaluza dimostrò che le due forze derivano da una versione pentadimensionale della relatività generale.

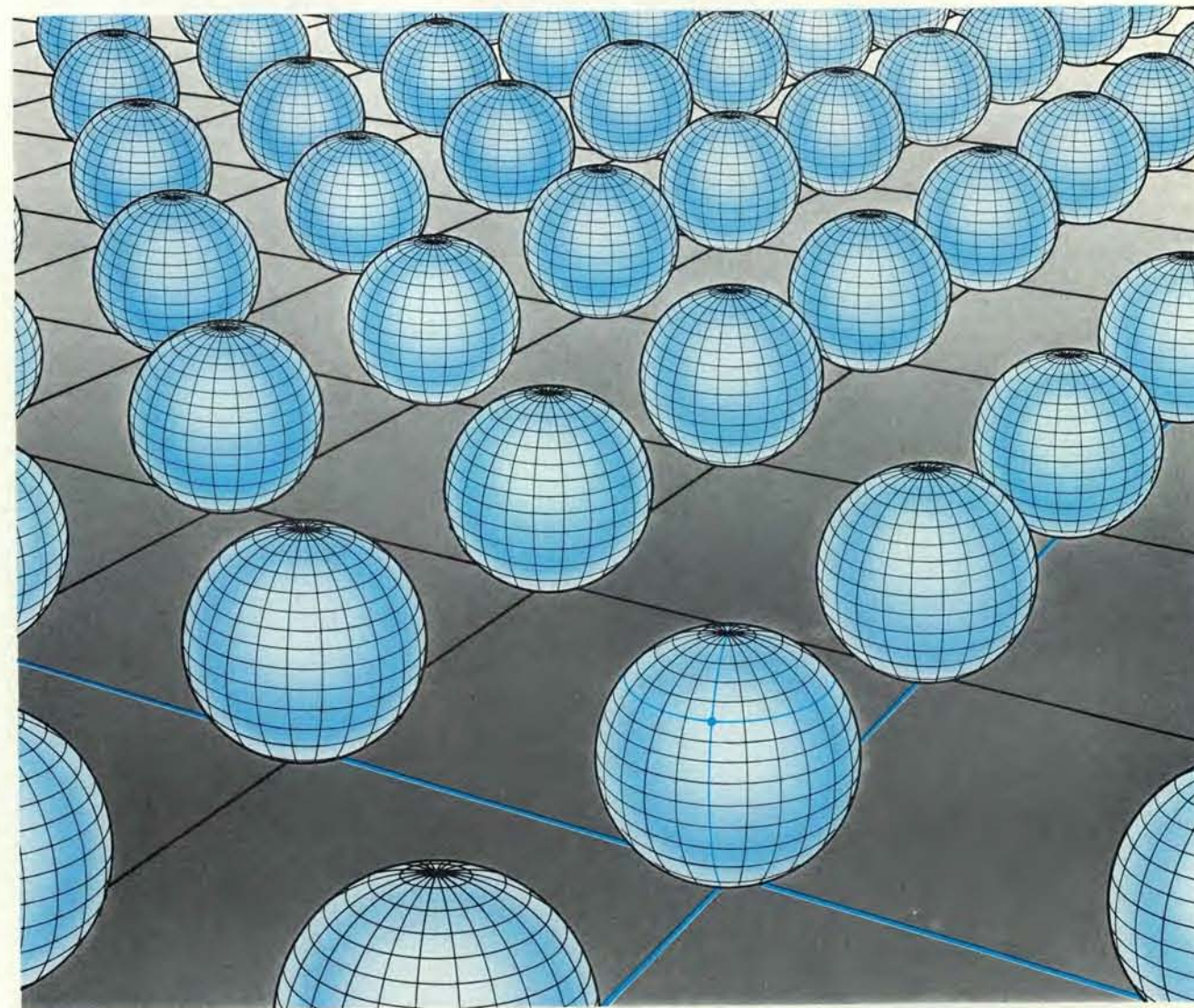
Nell'ultimo decennio molti fisici hanno prestato un rinnovato interesse al «programma» geometrico proposto da Kaluza per l'unificazione delle forze della natura. Nel programma attuale, però, vanno considerate strutture geometriche anche con più di cinque dimensioni, perché si conoscono quattro forze anziché due. Le due forze in più sono la forza nucleare forte, che lega tra loro protoni e neutroni all'interno del nucleo atomico, e la forza nucleare debole, responsabile di certi tipi di decadimento radioattivo. Inoltre, si è oggi accertato che non è possibile escludere da qualsiasi schema di unificazione gli effetti quantomeccanici. Uno fra i più attraenti sviluppi del programma attuale è una versione della teoria chiamata supergravità la quale, pur ammettendo diverse possibilità riguardo al numero di dimensioni dello spazio-tempo, presenta la massima eleganza matematica quando viene formulata in 11 dimensioni.

Perché sono necessarie 11 dimensioni? Questo numero deriva da una curiosa coincidenza matematica. Le teorie della supergravità si possono formulare con un numero qualsiasi di dimensioni dello spazio-tempo fino a 11, mentre con 12 o più dimensioni sembra che la teoria non sia più valida. D'altra parte, il numero minimo di dimensioni nascoste necessarie per sistemare le tre forze non gravitazionali in una teoria come quella di Kaluza è sette. Prese insieme alle quattro dimensioni dello spazio-tempo comune, le sette dimensioni nascoste porterebbero a un universo a 11 dimensioni. È importante notare che i requisiti matematici della supergravità coincidono con i limiti fisici imposti dalla descrizione delle forze.

La relatività generale

La teoria generale della relatività di Einstein è il coronamento delle ricerche della fisica classica. La supergravità, come qualsiasi altra teoria che si basi sulle idee geometriche di Kaluza per l'unificazione delle forze della natura, è essenzialmente un'estensione dei concetti della relatività generale. Einstein propose la relatività generale dopo nove anni di ricerche su una teoria della gravitazione: la teoria cercata doveva essere in accordo con la sua teoria della relatività ristretta e inoltre con l'osservazione sperimentale, che risaliva a Galileo, secondo la quale, in un campo gravitazionale, tutti i corpi seguono la stessa traiettoria. Einstein era convinto che, dal momento che la traiettoria di un corpo in caduta libera non dipende dalla sua massa né dalla sua composizione interna, il moto del corpo sotto l'azione della gravità deve essere correlato alle proprietà dello spazio-tempo stesso. Einstein indicò poi il modo di interpretare la forza come una manifestazione di una proprietà dello spazio-tempo chiamata curvatura.

Per comprendere meglio questa interpretazione, si consideri la superficie curva di una sfera. La superficie è bidimensionale perché sono necessarie due coordinate, come la latitudine e la longitudine, per individuare la posizione di un punto. La linea più breve che unisce due punti della sfera e che giace interamente sulla superficie è l'arco minore del cerchio massimo che passa per i due punti. (Questa proprietà geometrica fondamentale si applica comunemente nella scelta delle rotte aeree più convenienti.) Si può anche immaginare una superficie increspata più complessa della sfera, ma pure in questo caso esiste sulla superficie una linea di lunghezza minima che unisce due punti qualsiasi. Questa distanza viene chiamata geodetica.

Nella relatività generale lo spazio-tempo è l'analogo tetradimensionale di una superficie increspata perché sono necessarie quattro coordinate per individuare la posizione di un punto. Un punto dello spazio-tempo può essere un evento fisico, come la collisione tra due particelle, ed esso viene individuato precisando dove e quando accade, ossia per mezzo delle sue tre coordinate spaziali e del suo tempo. Una geodetica nello spazio-tempo è l'analogo di una geodetica su una superficie: è una linea nello spa-

Le sette dimensioni nascoste dell'universo, proposte in una teoria che cerca di unificare le forze della natura, possono essere descritte come una piccola struttura compatta quale una sfera associata a ogni punto dello spazio e a ogni istante del tempo. Nella teoria della relatività generale di Einstein lo spazio e il tempo sono riuniti in una struttura tetradimensionale chiamata spazio-tempo. Le osservazioni astronomiche mostrano che, su grande scala, lo spazio-tempo ha una geometria quasi piatta, o euclidea. Il piano raffigurato nell'illustrazione rappresenta la geometria dello spazio-tempo comune; le coordinate lungo un asse rappresentano lo spazio, mentre le coordinate lungo il secondo asse di riferimento rappresentano il tempo. Le sfere costruite alle intersezioni delle coordinate rappresentano le dimensioni nascoste e arrotolate che sono postulate nella nuova teoria. L'illustrazione può essere solo allusiva della struttura proposta, e le sfere dovrebbero essere immaginate tangenti al piano in ogni loro punto; inoltre, le sfere e il piano danno in realtà origine a solo quattro e non a 11 dimensioni. Le quattro dimensioni sono le quattro coordinate delle quali si devono fornire i valori per individuare la posizione di un punto *(in colore)*.

zio-tempo tra due eventi determinata dalla geometria dello spazio-tempo. Secondo la relatività generale, qualsiasi particella sulla quale agisce solo la forza di gravità segue una geodetica nello spazio-tempo; la relatività generale spiega così l'osservazione compiuta da Galileo secondo la quale tutti i corpi in caduta libera seguono una traiettoria comune.

La teoria unificata di Kaluza

Dal momento che la descrizione delle forze unificate da lui fatta aveva la stessa impostazione della relatività generale, Kaluza inviò il proprio saggio a Einstein per un consiglio. A quel tempo era possibile pubblicare un saggio soltanto se era stato avallato da un fisico ben conosciuto e inoltre, nella sua posizione di libero docente, Kaluza era poco autorevole e poteva disporre unicamente dei modesti proventi degli onorari versatigli dagli studenti che frequentavano le sue lezioni. Einstein, che aveva anch'egli iniziato la carriera come libero docente, fu subito affascinato dal saggio, ma in una serie di lettere inviate a Kaluza gli suggeriva di approfondire ulteriormente alcuni problemi della teoria prima della pubblicazione. Due anni e mezzo più tardi Einstein cambiò idea e inviò a Kaluza una cartolina nella quale gli comunicava l'intenzione di appoggiare la pubblicazione. L'articolo apparve nel 1921 nella rivista «Sitzungsberichte der Berliner Akademie» con il titolo *Il problema dell'unificazione in fisica*.

La ricerca di una descrizione unificata di tutti i fenomeni fisici apparentemente non correlati è sempre stata un tema di enorme importanza nell'indagine scientifica. Come abbiamo già detto, nella teoria di Kaluza forze comuni come la gravità e l'elettromagnetismo derivano da una versione pentadimensionale della relatività generale. Per spiegare il fatto che le cinque dimensioni non si osservano, Kaluza ipotizzò semplicemente che grandezze quali la curvatura non dipendono dalla quinta coordinata: le particelle seguono la geodetica nelle cinque dimensioni, ma le loro traiettorie appaiono a quattro dimensioni come quelle di particelle soggette all'azione combinata della forza di gravità e dell'elettromagnetismo.

Secondo il punto di vista attuale la più grave mancanza della teoria di Kaluza è che la gravità e l'elettromagnetismo non sono le sole forze fondamentali della natura. Nel 1919 la forza nucleare forte e la forza nucleare debole non erano ancora state scoperte perché il loro breve raggio d'azione è paragonabile al diametro del nucleo, e non erano ancora stati costruiti gli acceleratori capaci di verificare i processi dinamici a distanze così brevi.

All'epoca della pubblicazione dell'articolo di Kaluza la teoria presentava però un difetto ben più evidente: essa trascurava una serie di importanti fenomeni oggi conosciuti come effetti quantomeccanici. Kaluza era consapevole di questa mancanza e al termine del suo saggio scrisse: «Ogni [teoria classica, o deterministica e meccanicistica] che pretende di avere validità universale è minacciata dalla sfinge della fisica moderna, la teoria quantistica.» Ciononostante, nella teoria di Kaluza come nella teoria della relatività generale di Einstein è data per scontata una visione classica del mondo.

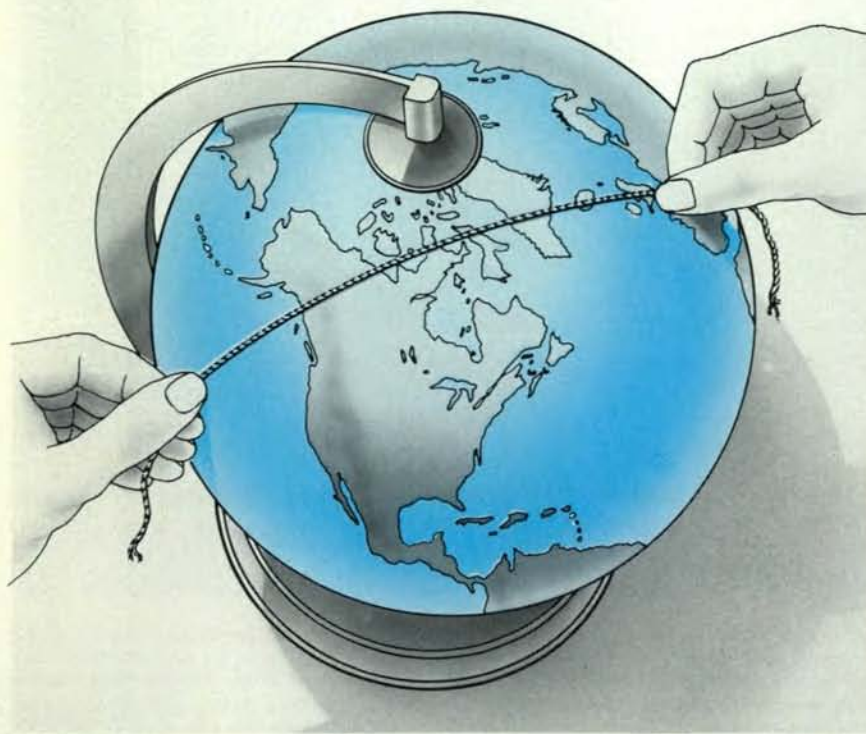
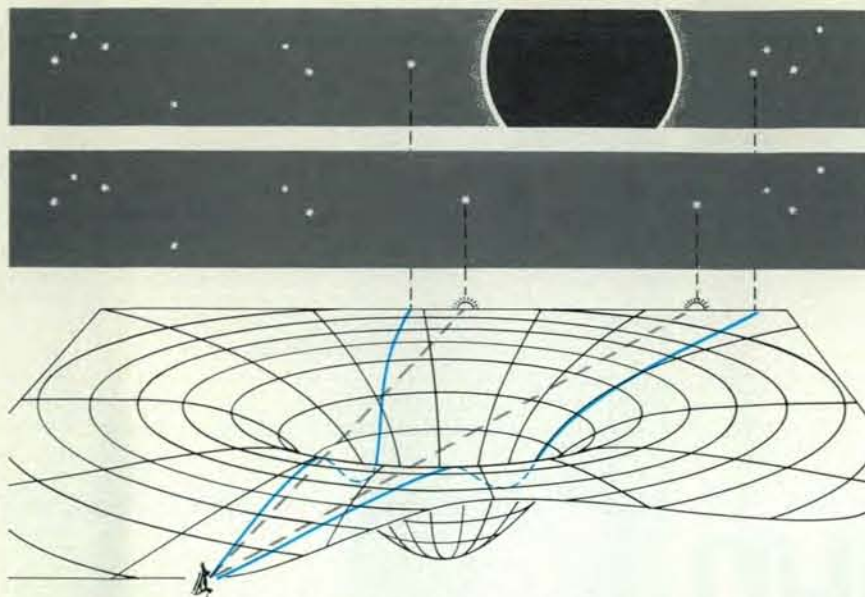
Secondo la concezione classica, tutti gli oggetti fisici - e con essi le più piccole particelle elementari - si comportano come proiettili sottoposti a una o più forze fondamentali. Per fenomeni di grande scala la concezione classica va abbastanza bene, mentre è del tutto incapace di spiegare processi in scala atomica. Nel 1919 erano già stati evidenziati molti dei difetti presenti nelle spiegazioni classiche dei processi atomici e subatomici.

Storicamente, la più grave carenza della fisica classica era la sua incapacità di spiegare lo spettro atomico. Gli esperimenti dimostrano che gli atomi emettono luce in righe spettrali discrete, corrispondenti a un insieme di frequenze, o colori, che sono caratteristiche dell'atomo emittente. Tuttavia, secondo la teoria classica, un atomo dovrebbe emettere luce di tutte le frequenze in quanto gli elettroni orbitanti di un atomo devono muoversi continuamente con moto a spirale verso il nucleo. Inoltre, nella descrizione classica il percorso a spirale degli elettroni condurrebbe rapidamente al collasso dell'atomo e quindi la materia come noi la conosciamo non potrebbe esistere.

La necessità di risolvere questo enigma e altre difficoltà portarono allo sviluppo della meccanica quantistica, nella quale si abbandona il rigoroso determinismo della teoria classica e le traiettorie a spirale degli elettroni attorno al nucleo vengono quindi sostituite da configurazioni ondulatorie nello spazio-tempo: l'intensità di una configurazione ondulatoria determina la probabilità di trovare un elettrone in un particolare punto.

Onde stazionarie corrispondono a stati di moto a lunga vita dell'elettrone e ogni stato di moto possiede un'energia caratteristica. La luce viene emessa a frequenze discrete che corrispondono a righe spettrali discrete quando l'elettrone salta improvvisamente da uno stato a un altro. Lo stato di moto corrispondente alla minima energia permessa è stabile e quindi nella teoria quantistica gli atomi non collassano come avverrebbe in base alla teoria classica. Le configurazioni ondulatorie degli elettroni sono soluzioni di un'equazione differenziale formulata da Erwin Schrödinger, nella quale compaiono come variabili sia il tempo sia le tre coordinate spaziali.

La quinta dimensione

Nel 1926, ancora agli albori dell'era quantistica, il fisico svedese Oskar Klein si prefisse di stabilire se la meccanica quantistica era o meno compatibile con

La geodetica è la distanza più breve che intercorre tra due punti che siano situati sulla superficie di una sfera. Su un mappamondo la geodetica è l'arco minore del cerchio massimo che passa per quei due punti. È possibile determinarla mediante una cordicella tesa tra i due punti.

La curvatura della luce stellare in prossimità del Sole è un effetto previsto dalla teoria della relatività generale di Einstein. Secondo la teoria, nelle vicinanze del Sole la struttura geometrica dello spazio-tempo viene curvata dalla massa solare nel modo suggerito dall'insieme di assi coordinati curvi rappresentati nel grafico. La luce deve seguire una geodetica nello spazio-tempo, quindi le linee di vista verso le stelle vengono curvate quando queste sono vicine al disco solare *(in colore)*. Se si osservano durante un'eclisse, le stelle appaiono spostate dal Sole. Le linee in nero tratteggiate indicano le linee di vista quando il Sole non è molto vicino.

la teoria pentadimensionale di Kaluza. Klein formulò una versione dell'equazione di Schrödinger con cinque anziché quattro variabili e dimostrò che le soluzioni si possono interpretare come onde che si muovono in campi gravitazionali ed elettromagnetici dello spazio-tempo comune a quattro dimensioni. (Nella meccanica quantistica si interpretano le onde anche come particelle.) Oggi si chiamano di Kaluza-Klein tutte le teorie che tentano, secondo uno schema quantomeccanico, di unificare le forze fondamentali della natura in uno spazio-tempo con più di quattro dimensioni.

Nei saggi originali di Kaluza e di Klein non è chiaro se la quinta dimensione va intesa come una realtà fisica o semplicemente come un artificio matematico necessario per ricavare la gravità e l'elettromagnetismo in modo coerente. L'introduzione della meccanica quantistica suggerisce però risposte attendibili a numerosi e importanti interrogativi sulla realtà fisica di una dimensione in più. In che senso la nuova dimensione potrebbe essere una realtà fisica? Perché non è stato scoperto finora un aspetto così fondamentale dell'universo? Come si potrebbe scoprire sperimentalmente la dimensione in più?

Per cominciare a rispondere, si consideri una retta di lunghezza indefinita a ogni punto della quale sia associato un piccolo cerchio. Se si costruisce effettivamente un cerchio in ogni punto della retta, la struttura risultante è un cilindro di lunghezza indefinita: si può dire che la retta e il cerchio unidimensionali generano il cilindro bidimensionale.

In modo analogo si può generare una struttura tetradimensionale dal piano bidimensionale e dalla sfera bidimensionale. Si può pensare la nuova struttura come un piano in ogni punto del quale viene costruita una sfera: è tetradimensionale perché sono necessarie due coordinate per individuare la posizione di un punto nel piano e altre due per individuare un punto sulla sfera (*si veda l'illustrazione a pagina 95*).

La retta e il piano dei due esempi precedenti rappresentano la geometria quasi piatta dello spazio-tempo tetradimensionale nel quale viviamo, mentre il cerchio e la superficie sferica rappresentano la dimensione o le dimensioni in più di uno spazio-tempo con un maggior numero di dimensioni. Uno spazio-tempo pentadimensionale si può intendere come la struttura generata da un cerchio e da un comune spazio-tempo tetradimensionale; una possibile struttura di uno spazio-tempo esadimensionale si genera con lo spazio-tempo comune e con la superficie di una sfera. In queste strutture a ogni punto dello spazio e a ogni istante di tempo sono associati un cerchio o una sfera.

Ora siamo in grado di spiegare come, nella teoria di Kaluza, la quinta dimensione dello spazio-tempo possa essere reale, anche se fino a oggi non è ancora stata scoperta. Un concetto fondamentale della meccanica quantistica è il principio di indeterminazione di Werner Heisenberg. Qualsiasi particella può essere interpretata come un pacchetto di onde diffuse in una certa regione di spazio e, in base al principio di indeterminazione, le dimensioni minime della regione dipendono dall'energia della particella: maggiore è l'energia della particella, minori sono le dimensioni minime della regione.

Per rivelare una piccola struttura spaziale si deve usare un microscopio, ossia uno strumento che «illumina» una struttura con fotoni di luce, elettroni o fasci di qualche altra particella. La risoluzione del microscopio è la dimensione minima della regione che si può illuminare, e quindi, secondo il principio di indeterminazione, la risoluzione dipende dall'energia delle particelle del fascio incidente; ne risulta che per poter osservare strutture sempre più piccole sono necessarie particelle di energia sempre più elevata.

Supponiamo che la quinta dimensione sia arrotolata in un cerchio estremamente piccolo: per rivelarlo, l'energia delle particelle che lo illuminano dovrebbe essere sufficientemente elevata; particelle con energia troppo bassa finirebbero infatti con il distribuirsi uniformemente sul cerchio ed esso non potrebbe essere rivelato. I più potenti acceleratori attuali producono particelle la cui energia è sufficientemente alta da risolvere strutture con un diametro di anche 10^{-16} centimetri; se nella quinta dimensione il cerchio è più piccolo di 10^{-16} centimetri, potrebbe non essere stato finora risolto.

Particelle «massicce»

Esiste un modo più indiretto con cui si potrebbe dedurre l'esistenza di una quinta dimensione spaziale. Proprio come nell'atomo le configurazioni ondulatorie stazionarie corrispondono a stati di moto a lunga vita degli elettroni orbitanti, così le onde stazionarie sul cerchio della quinta dimensione corrispondono a particelle che si potrebbero osservare in laboratorio. Le configurazioni ondulatorie stazionarie devono adattarsi esattamente sulla circonferenza del cerchio; perciò o l'onda deve avere un'ampiezza costante oppure l'intero cerchio deve contenere un numero intero di oscillazioni: una, due o tre oscillazioni, e così via (*si veda l'illustrazione a pagina 101*).

La massa di ogni particella osservabile dipende dalla sua lunghezza d'onda, che è il rapporto tra la circonferenza del cerchio e il numero di oscillazioni che l'onda esegue attorno al cerchio: minore è la lunghezza d'onda, maggiore è l'energia dell'onda e più alta è la massa della particella associata. Nella teoria di Kaluza le particelle di massa minore sono quelle associate a lunghezze d'onda infinita; in altre parole, nella quinta dimensione l'ampiezza dell'onda è costante e le particelle hanno massa nulla.

Nella teoria la prima particella «massiccia» è quella la cui lunghezza d'onda è uguale alla circonferenza del cerchio; la sua massa è inversamente proporzionale alla circonferenza. La massa della seconda particella pesante è doppia della prima, perché corrisponde alla lunghezza d'onda contenuta esattamente due volte nella circonferenza del cerchio. Analogamente, le altre configurazioni ondulatorie stazionarie consentite sul cerchio generano una serie di particelle le cui masse sono multipli interi della massa della prima particella pesante.

Una argomentazione introdotta da Klein consente di stimare la massa della prima particella pesante. Dal momento che la teoria di Kaluza tenta di unificare la forza di gravità e l'elettromagnetismo, la prima particella pesante ha anche una carica elettrica che è inversamente proporzionale alla circonferenza del cerchio. D'altra parte, la carica di tutte le particelle elementari osservate è un multiplo intero della carica dell'elettrone, cosicché se si ipotizza che la prima particella pesante porti quella carica, se ne può calcolare la massa. La risposta è spaventosamente grande: la massa è 10^{16} volte quella del protone, che è a sua volta più pesante di 10 000 batteri. Né gli attuali né i futuri acceleratori potranno mai produrre tali particelle, che però potrebbero essere state prodotte nel big bang. Da allora la maggior parte di esse dovrebbe essere decaduta, ma alcune potrebbero essere ancora rivelabili.

Dal momento che le particelle massicce della teoria di Kaluza sono così pesanti, la sola particella della teoria che potrebbe corrispondere alle particelle attualmente osservate è quella di massa nulla. Oggi sappiamo, anche se la cosa non venne valutata nella dovuta misura all'epoca in cui fu formulata la teoria, che effetti quantomeccanici più sottili possono portare a una massa finita, non nulla, per la particella prevista dalla teoria. La particella priva di massa della teoria di Kaluza e altre particelle con massa nulla nelle generalizzazioni della teoria possono spiegare, almeno in linea di principio, le particelle osservate.

Anche la circonferenza del cerchio nella quinta dimensione che potrebbe dare origine alle particelle massicce previste dalla teoria è corrispondentemente piccola: circa 10^{-30} centimetri. Per risolvere una struttura di tali dimensioni con uno strumento basato sulle attuali tecnologie sarebbe necessario un acceleratore con un diametro di molti anni-luce.

Dopo le indagini di Klein e il successivo lavoro di Einstein e di Wolfgang Pauli, vi furono pochi progressi nell'idea fondamentale di unificazione di Kaluza fino alla fine degli anni settanta. Infatti, fino ad allora, la maggior parte delle ricerche sull'unificazione delle forze si basava su una strategia che non richiedeva uno spazio-tempo con un maggior numero di dimensioni. La strategia si può rintracciare in una proposta diversa di unificazione della gravità e dell'elettromagnetismo avanzata dal matematico tedesco Hermann Weyl nel 1918. L'idea centrale della teoria di Weyl è che la descrizione di una forza non viene alterata da una qualsivoglia modifica delle scale di lunghezza dei regoli o delle scale temporali degli orologi impiegati come strumenti di misura nei vari punti dello spazio-tempo. Questo principio è chiamato invarianza di gauge, da *gauge* (strumento di misura), al quale si riferiva Weyl. Una teoria di questo tipo è chiamata teoria di campo di gauge o, per brevità, teoria di gauge.

L'unificazione elettrodebole

La teoria originaria di Weyl non forniva una corretta interpretazione fisica della gravità ed è stata abbandonata. Ciononostante il principio di invarianza di gauge è diventato il perno delle moderne teorie sulle particelle elementari. Nel 1954 C. N. Yang, della State University of New York a Stony Brook, e Robert L. Mills, della Ohio State University, svilupparono una classe di teorie di gauge, le teorie di gauge non abeliane, che sono un'importante generalizzazione della teoria dell'elettromagnetismo di Maxwell in cui assume un ruolo centrale la teoria matematica dei gruppi di simmetria. Nella teoria dei gruppi si studiano operazioni, quali le rotazioni e le riflessioni speculari di oggetti solidi, che lasciano inalterato l'aspetto degli oggetti: per esempio, l'aspetto di una sfera non cambia dopo una qualsiasi rotazione rigida attorno al suo centro e il gruppo che esprime matematicamente questa simmetria è chiamato $SU(2)$.

Molti fisici teorici hanno studiato teorie di gauge non abeliane. Nel 1967 Steven Weinberg, attualmente all'Università del Texas ad Austin, Abdus Salam del Centro internazionale di fisica teorica di Trieste e John C. Ward, oggi alla Macquarie University del Nuovo Galles del Sud, applicarono alcuni importanti contributi di Peter Higgs dell'Università di Edimburgo, di Sheldon Lee Glashow della Harvard University e di altri per dimostrare che una teoria di gauge non abeliana potrebbe unificare la forza elettromagnetica e la forza nucleare debole. Alcune previsioni di questa teoria, chiamata teoria elettrodebole, sono state confermate sperimentalmente all'inizio degli anni settanta, ma la dimostrazione più spettacolare si è avuta nel 1983 al CERN, l'Organizzazione europea per la ricerca nucleare, allorché furono scoperte tre particelle, i bosoni vettori W^+, W^- e Z^0, aventi esattamente la massa prevista dalla teoria elettrodebole.

Il successo della teoria elettrodebole indusse i fisici teorici a proporre un'altra teoria di gauge non abeliana, chiamata cromodinamica quantistica, che può descrivere la forza nucleare forte. In questa teoria il protone e il neutrone sono formati da particelle ancor più elementari, i quark, e la forza forte deriva dalle interazioni dei quark con otto bosoni vettori chiamati gluoni; sembra che anche la cromodinamica quantistica sia confermata sperimentalmente.

21 aprile 1919

L'idea che le grandezze del campo elettrico siano mutilate... ha spesso tormentato anche me. L'idea, però, che questo si possa ottenere con un mondo cilindrico a cinque dimensioni non mi è mai venuta e sembrerebbe anche del tutto nuova. A prima vista mi piace moltissimo...

Se, leggendo la sua dettagliata esposizione, non sorgerà qualche grave obiezione, sarò lieto di presentare alla locale accademia questo suo saggio sull'argomento.

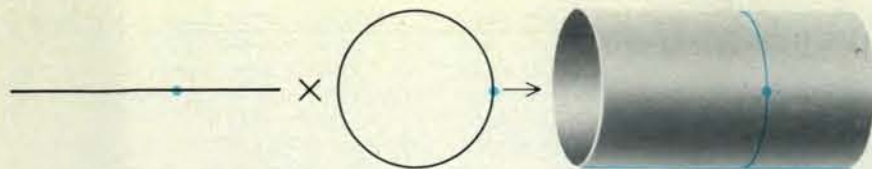
28 aprile 1919

Ho letto attentamente il saggio e l'ho trovato davvero interessante. Per ora non escludo la possibilità di appoggiarlo. D'altra parte, devo riconoscere che le argomentazioni finora esposte non appaiono sufficientemente convincenti. Le suggerirei di leggere quanto segue (magari prima di pubblicare il saggio anche se non mi piace di darle consigli in materia).

Seguendo la sua idea fondamentale, si dovrebbe supporre che le geodetiche che sono oblique alle sezioni... dovrebbero fornire le traiettorie di particelle elettricamente cariche sotto l'azione simultanea del campo gravitazionale e di quello elettrico. *Se lei fosse in grado dimostrare che questo accade con la precisione garantita dalle nostre conoscenze empiriche,* rimarrei pienamente convinto della sua teoria.

14 ottobre 1921

Sto avendo un ripensamento sul fatto di averla indotta a desistere, due anni fa, dalla pubblicazione della sua idea su un'unificazione della gravitazione e della elettricità. Il suo modo di affrontare il problema mi pare in ogni caso più valido di quello di H. [Hermann] Weyl. Se vuole presenterò al più presto il saggio all'accademia, purché me lo invii.

Le lettere di Einstein a Theodor Franz Eduard Kaluza mostrano come Einstein modificò il proprio atteggiamento rispetto alle idee di Kaluza. Le date indicano che passarono oltre due anni prima che Einstein appoggiasse la pubblicazione del saggio di Kaluza, che uscì nel 1921.

La teoria di Kaluza considera la quinta dimensione come un cerchio associato a ogni punto del comune spazio-tempo. Se di questo si rappresenta una sola dimensione con una retta, si può visualizzare l'analogo della struttura pentadimensionale proposta da Kaluza: una retta a ogni punto della quale è associato un cerchio, o, in altre parole, un cilindro. Una sezione circolare trasversale del cilindro rappresenta la struttura dello spazio-tempo pentadimensionale vuoto.

Pur essendo la teoria elettrodebole e la cromodinamica quantistica teorie di gauge alquanto differenti, le tre forze da esse descritte si possono ulteriormente unificare in un'unica teoria di gauge non abeliana basata su un più ampio gruppo matematico di simmetria. Tali teorie si chiamano teorie di grande unificazione; le loro previsioni non sono state ancora confermate sperimentalmente, ma i concetti sono talmente attraenti che molti fisici sono convinti che qualche loro versione riuscirà a fornire una corretta spiegazione unificata delle forze forte, debole ed elettromagnetica.

Quella che manca nelle teorie di grande unificazione è la forza di gravità ed è perciò naturale chiedersi se queste teorie possano essere fuse insieme con la gravitazione come una teoria di Kaluza-Klein con un maggior numero di dimensioni. La teoria originale di Kaluza richiedeva cinque dimensioni perché essa comprendeva soltanto un bosone vettore, cioè il fotone associato alla forza elettromagnetica. La forza nucleare debole richiede i tre bosoni vettori scoperti di recente, la forza nucleare forte gli otto gluoni, mentre la grande unificazione richiede da 10 a 500 altri bosoni vettori. Il numero esatto di bosoni vettori addizionali dipende da quale versione della teoria di grande unificazione si adotta.

Le moderne teorie di Kaluza-Klein

Anche se non esiste una corrispondenza biunivoca tra il numero di bosoni vettori necessari e il numero di dimensioni, è approssimativamente corretto affermare che più bosoni vettori richiedono più dimensioni dello spazio-tempo; di conseguenza l'inclusione delle forze forte e debole nello schema teorico di Kaluza-Klein richiederebbe uno spazio-tempo con ancor più di cinque dimensioni. Le dimensioni in più potrebbero essere fisicamente reali e tuttavia inosservate a condizione che esse si arrotolino in una «superficie» con un maggior numero di dimensioni analoga al cerchio della teoria di Kaluza o alla superficie di una sfera.

I recenti tentativi di inserire le forze forte e debole in una teoria di Kaluza-Klein hanno avuto inizio con le ricerche di Bryce S. DeWitt dell'Università del Texas ad Austin, di Y. M. Cho della Università nazionale di Seoul, di Peter G. O. Freund e Mark A. Rubin dell'Università di Chicago, di Eugene Cremmer, Bernard Julia e dello scomparso Joel Scherk dell'Università di Parigi e di John H. Schwartz del California Institute of Technology.

Il primo problema che si presenta per le moderne teorie di Kaluza-Klein è il numero di dimensioni addizionali da inserire. Non essendovi ancora un consenso generale su quale sia la versione corretta delle teorie di grande unificazione, anche il numero di bosoni vettori è incerto. Pertanto il numero di dimensioni addizionali in una teoria di Kaluza-Klein è sia incerto sia arbitrario.

Il secondo problema è quello di spiegare le particelle elementari osservate. In teorie quantistiche come le teorie di gauge non abeliane vi sono due classi di particelle elementari, cioè i bosoni e i fermioni. Abbiamo già parlato dei bosoni, che sono i portatori delle forze fondamentali; per esempio, nella visione quantomeccanica, la forza di gravità è causata da uno scambio continuo di bosoni chiamati gravitoni tra due corpi dotati di massa. Il risultato dello scambio si manifesta, in laboratorio, come un'attrazione tra i due corpi. Non vi sono difficoltà nel ricavare i bosoni da una teoria di Kaluza-Klein. Il campo gravitazionale con un maggior numero di dimensioni può facilmente condurre ai bosoni del mondo tetradimensionale.

I fermioni, che rappresentano la seconda classe di particelle elementari, hanno in fisica un ruolo completamente diverso in quanto, a differenza dei bosoni, che trasmettono le forze, costituiscono tutta la materia ponderabile dell'universo: l'elettrone, il neutrone, il protone e il neutrino sono fermioni, come pure sono fermioni gli stessi quark che costituiscono il neutrone e il protone.

Come si possono spiegare i fermioni in una teoria di Kaluza-Klein? Non si possono ricavare da un campo gravitazionale bosonico; il solo modo per ottenerli è quello di aggiungere uno o più campi fermionici alla teoria con un più elevato numero di dimensioni: i campi porterebbero in tal caso ai fermioni osservati nelle quattro dimensioni. Il numero di campi fermionici inclusi nella teoria è arbitrario perché non esiste alcun fondamento teorico su cui basarsi.

La supergravità

Sono molti gli studi interessanti su teorie di Kaluza-Klein con un numero di dimensioni arbitrario dove i campi fermionici vengono aggiunti «a mano»; tuttavia l'arbitrarietà allontana dalla semplicità dell'idea originale di Kaluza. È auspicabile una teoria nella quale il numero di campi fermionici e il numero di dimensioni siano fissati naturalmente dalla struttura della teoria.

Una teoria siffatta è la supergravità. In primo luogo si tratta di un'estensione della relatività generale nella quale bosoni e fermioni sono trattati su un piede di parità: il gravitone bosonico, per esempio, ha un partner fermionico chiamato gravitino. Nella versione della relatività generale di Einstein si possono aggiungere o togliere fermioni a volontà, mentre nella supergravità esiste un partner fermionico per ogni bosone; i fermioni necessari, quindi, per descrivere la struttura della materia sono presenti nella teoria fin dall'inizio.

Nella supergravità anche il numero di dimensioni è fisso. Come abbiamo detto sopra, le teorie della supergravità probabilmente non sono valide per un numero di dimensioni superiore a 11. Oltre tale numero non è possibile trovare i requisiti matematici per una correlazione tra campi bosonici e campi fermionici. Inoltre, Edward Witten della Princeton University ha dimostrato che alle quattro dimensioni dello spazio-tempo si devono aggiungere almeno sette dimensioni nascoste per poter includere in uno schema di Kaluza-Klein le forze forte, debole ed elettromagnetica. Esiste un terzo aspetto della supergravità a 11 dimensioni che è circostanziale, ma di alto interesse teorico: mentre per un numero di dimensioni inferiore a 11 esistono parecchie versioni della supergravità matematicamente distinte, a 11 dimensioni la teoria è unica.

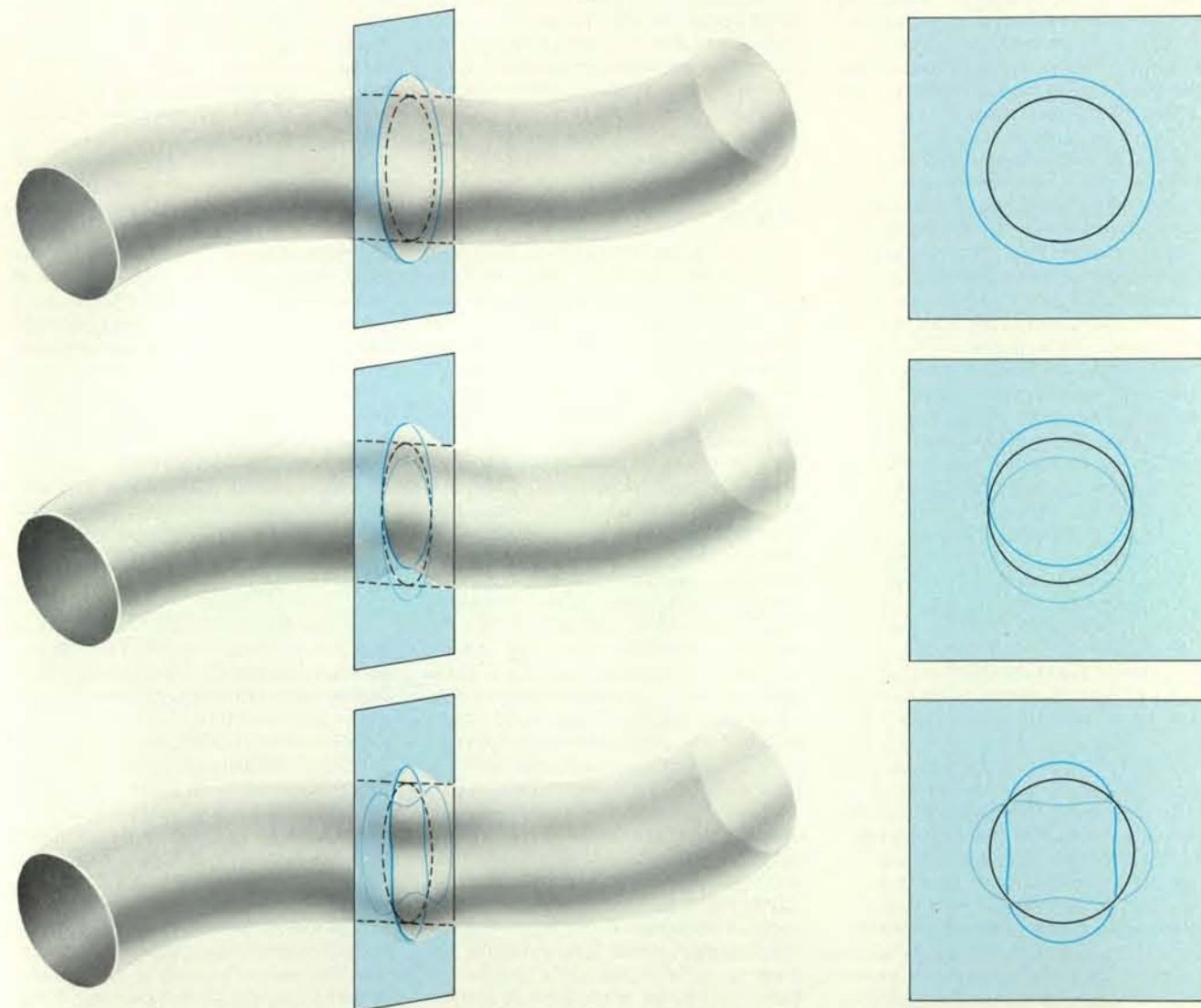
Gli «ingredienti» minimi di una teoria di Kaluza-Klein comprendono il campo gravitazionale, che dà origine ai bosoni, e un campo fermionico, che spiega i fermioni del nostro mondo. Oltre al campo gravitazionale, deve esserci almeno un altro campo bosonico, il quale funga da sorgente che compatta, o arrotola, le dimensioni addizionali nascoste. È notevole che la versione a 11 dimensioni della supergravità contenga esattamente tutti e tre questi ingredienti.

Per il teorico risulta ancor più sorprendente il fatto che il campo bosonico in più conduca naturalmente a solo due tipi di compattazione. In un tipo sette delle 11 dimensioni si arrotolano in una piccola struttura nascosta: tale compattazione spiegherebbe perché il numero di dimensioni facilmente osservabili nel mondo è quattro. L'alternativa è che si arrotolino soltanto quattro dimensioni e questo scenario condurrebbe a un mondo a sette dimensioni. Può darsi che i futuri fisici scopriranno perché si preferisce un mondo a quattro dimensioni.

Per mettere a punto una teoria di Kaluza-Klein della supergravità a 11 dimensioni i fisici devono prima risolvere le equazioni della supergravità. Molte soluzioni danno origine a una struttura dello spazio-tempo generata da uno spazio-tempo tetradimensionale e da una piccola superficie chiusa a sette dimensioni. A questo punto si studia il gruppo di simmetria di ciascuna superficie corrispondente a una soluzione delle equazioni e tale gruppo determina la teoria di gauge non abeliana che si deve unificare alla gravità. Superfici chiuse differenti hanno gruppi di simmetria differenti e ciascuno determina una differente teoria di grande unificazione delle forze non gravitazionali.

La fase finale nello sviluppo di una teoria di Kaluza-Klein è l'analisi delle configurazioni ondulatorie stazionarie complesse, che sono consentite dalle superfici chiuse e determinano le masse e le altre proprietà delle particelle previste dalla teoria nel comune spazio-tempo tetradimensionale. Ognuna delle superfici a sette dimensioni che appaiono come soluzione delle equazioni della supergravità va analizzata in questo modo.

Risultati teorici

Nella maggior parte degli studi sono stati presi in considerazione due casi. Nel primo caso le dimensioni arrotolate formano la struttura a sette dimensioni più semplice e più simmetrica possibile, cioè l'analogo a sette dimensioni della sfera.

Nella teoria di Kaluza, le particelle possono essere associate alla quinta dimensione circolare arrotolata. Secondo la meccanica quantistica, ogni particella può essere anche considerata come un'onda; per questo motivo, se qualche multiplo intero della lunghezza d'onda è contenuto esattamente nella circonferenza del cerchio nella quinta dimensione, la particella corrispondente alla lunghezza d'onda dovrebbe poter esistere nel comune spazio-tempo tetradimensionale. Il primo tipo di onda che si adatta al cerchio è un'onda di ampiezza costante estesa all'intera circonferenza. Se si rappresenta il comune spazio-tempo curvo con una linea curva, lo spazio-tempo con un maggior numero di dimensioni generato dalla linea curva e dalla quinta dimensione circolare è un cilindro curvo. L'onda di ampiezza costante appare come un rigonfiamento del cilindro (*in alto*); in alto a destra è illustrata la sua sezione trasversale. Nella teoria di Kaluza la particella corrispondente a questa onda è priva di massa. Il secondo tipo di onda compie un'oscillazione attorno al cerchio e appare anch'essa come un rigonfiamento, mentre la sua sezione trasversale è in realtà un'onda cosinusoidale tracciata attorno al cerchio come se il cerchio fosse l'asse orizzontale di un grafico (*al centro*). La figura chiusa formata dall'onda è indicata dalla curva in colore intenso; la figura precede, o ruota, attorno al cerchio, e una delle sue successive orientazioni è mostrata dalla curva in colore tenue. La teoria di Kaluza prevede che la massa della particella associata a quest'onda sia 10^{16} volte quella del protone. Il terzo tipo di onda e quelli di ordine superiore dividono il cerchio in due, tre o più parti uguali. Il terzo tipo di onda è un'onda cosinusoidale la cui lunghezza d'onda è contenuta esattamente due volte nel cerchio (*in basso*); anch'essa precede nel modo indicato e la particella a essa associata ha una massa doppia di quella associata alla seconda onda.

In gran parte il lavoro sulla sfera a sette dimensioni è stato svolto da Michael J. Duff e Christopher N. Pope dell'Imperial College of Science and Technology di Londra, da François Englert dell'Université Libre di Bruxelles, da Bernard de Wit dell'Università statale di Utrecht e da Hermann Nicolai del CERN.

Il secondo caso è un insieme di superfici aventi il gruppo di simmetria necessario per le forze forte, debole ed elettromagnetica. Queste superfici sono state studiate da Witten, da Leonardo Castellani, Riccardo D'Auria e Pietro Fré dell'Università di Torino e da altri.

Sfortunatamente i risultati particolareggiati degli studi non prevedono un mondo a quattro dimensioni che assomigli a quello che conosciamo. Vi sono al riguardo tre importanti problemi. Il primo è quello della chiralità perché riguarda l'orientamento destrorso o sinistrorso dei fermioni previsto dalla teoria. (La chiralità di un fermione è determinata dal verso del suo spin quantomeccanico rispetto alla direzione del suo moto.) Tutte le strutture a 11 dimensioni studiate finora prevedono lo stesso numero di neutrini sinistrorsi e destrorsi, mentre i neutrini osservati in natura sono sempre sinistrorsi; pare quindi che non esistano neutrini destrorsi.

Il secondo è il problema cosmologico e riguarda la curvatura prevista per il comune spazio-tempo tetradimensionale. Se si avanza la ragionevole ipotesi che le sette dimensioni addizionali formino una struttura compatta talmente piccola da non essere ancora stata osservata, le restanti quattro dimensioni dello spazio-tempo acquistano un'elevata curvatura. Ciò è in contrasto con le osservazioni astronomiche, in base alle quali la curvatura dell'universo su grande scala è nulla o quasi nulla. Nelle teorie di Kaluza-Klein non basate sulla supergravità il problema si può evitare introducendo nelle equazioni una costante, detta costante cosmologica, il cui effetto è quello di cancellare la curvatura dello spazio-tempo tetradimensionale anche quando le altre sette dimensioni sono altamente compatte. Tale possibilità di adattare le equazioni fondamentali non esiste nella supergravità a 11 dimensioni.

Il terzo problema della supergravità a 11 dimensioni è il problema quantistico, ma si spera che la sua soluzione possa eliminare anche i due problemi precedenti. Le teorie alla base del programma di Kaluza-Klein sono fondate su equazioni quantomeccaniche; queste portano a grandezze infinite che non hanno alcuna ovvia interpretazione fisica. Le grandezze infinite presentano una difficoltà generale per quasi tutte le teorie quantistiche della gravità e per evitarle i teorici sono stati costretti a eseguire approssimazioni che trascurano alcuni degli effetti quantistici. Alla fine si può sperare o di dimostrare che gli infiniti sono dovuti alla procedura di approssimazione e non alla teoria stessa, oppure di trovare una teoria particolare nella quale gli infiniti siano assenti.

Negli ultimi mesi alcuni fisici teorici si sono entusiasmati alla prospettiva che il problema delle grandezze infinite e forse gli altri problemi che abbiamo citato si possano risolvere con un tipo di teoria chiamato «teoria della supercorda» (si veda l'articolo *Modelli a risonanza duale delle particelle elementari* di John H. Schwarz in «Le Scienze» n. 82, giugno 1975). Le teorie della supercorda presentano alcune attraenti proprietà tipiche della supergravità. Per poter essere matematicamente coerenti, esse devono essere costruite nello spazio-tempo a 10 dimensioni, e a 10 dimensioni sono poche le teorie possibili. Per un certo tempo si è creduto che le grandezze infinite fossero assenti nella teoria della supergravità al primo livello di approssimazione degli effetti quantistici. Oggi alcuni fisici pensano che esse siano assenti a ogni livello di approssimazione.

La teoria della supercorda

In una teoria della corda le particelle sono associate ai moti vibrazionali di una corda unidimensionale in uno spazio con un maggior numero di dimensioni. La differenza fondamentale tra una teoria della corda e una teoria di campo, quale la supergravità, sta nel modo in cui si deve contare il numero di particelle previsto dalle due teorie. Se le sette dimensioni in più della supergravità con un maggior numero di dimensioni non fossero arrotolate in una superficie chiusa, la supergravità a 11 dimensioni senza compattazione prevederebbe un numero di particelle finito. Un numero infinito di particelle nasce nella supergravità solo a causa della compattazione. Per esempio, nella teoria pentadimensionale di Kaluza c'è una serie infinita di particelle perché c'è una serie infinita di configurazioni ondulatorie stazionarie che si adattano alla quinta dimensione circolare. D'altra parte, nella teoria della supercorda vi è un numero infinito di particelle anche senza compattazione delle dimensioni in più. Il numero infinito di particelle della teoria della supercorda corrisponde al numero infinito di configurazioni ondulatorie che possono persistere sulla corda.

La maggior parte delle particelle che sorgono nella teoria della supercorda hanno una massa estremamente grande: più di 10^{19} volte la massa del protone; ciononostante, la teoria prevede anche circa 1000 particelle prive di massa. Fino a poco tempo fa si pensava che le mutue interazioni di queste particelle fossero equivalenti alle interazioni descritte da una versione della supergravità a 10 dimensioni, e vi erano due motivi per non studiare a fondo questa versione. In primo luogo, pareva non esistessero soluzioni alle equazioni della teoria nella quale sei dimensioni si arrotolano e lasciano uno spazio-tempo tetradimensionale con proprietà «ragionevoli». In secondo luogo, le equazioni stesse diventano incoerenti quando vengono interpretate a livello quantistico. La versione della supergravità a 10 dimensioni, e quindi le mutue interazioni delle particelle prive di massa descritte dalla teoria della supercorda, non apparivano interessanti per il programma Kaluza-Klein.

Di recente Michael Green del Queen Mary College di Londra e Schwarz hanno dimostrato che le interazioni delle particelle prive di massa della teoria della supercorda differiscono leggermente dalle loro interazioni nella versione della supergravità a 10 dimensioni. Gli effetti sono sottili e sono dovuti al numero infinito di particelle pesanti presenti nella teoria della supercorda, ma non nella supergravità senza compattazione. Se si tiene conto degli effetti delle particelle pesanti, si ottengono equazioni coerenti a livello quantistico.

Questo recente successo ha stimolato un rinnovato e vigoroso sforzo per compattare le sei dimensioni in più della teoria della supercorda. Per molti aspetti il problema è ancor più difficile che non nella supergravità a 11 dimensioni in quanto le proprietà delle superfici a sei dimensioni richieste nella teoria della supercorda sono matematicamente più complesse delle proprietà, per esempio, della sfera a sette dimensioni. Ciononostante, vi sono molti stimoli alla soluzione del problema e vi sono indicazioni che gli altri due problemi importanti della supergravità, cioè il problema della chiralità e il problema cosmologico, non sorgono nella teoria della supercorda.

Futuri sviluppi

Spesso trascorre molto tempo tra lo sviluppo di eleganti concetti teorici e la precisa formulazione di previsioni verificabili sperimentalmente. Sono stati necessari, per esempio, 13 anni per trovare il modo corretto di applicare le teorie di gauge non abeliane all'unificazione delle forze fondamentali. L'attuale mancanza di chiare indicazioni sulla correttezza sperimentale delle idee della supergravità e della teoria di Kaluza-Klein non significa necessariamente che esse siano errate, e può darsi che sia semplicemente necessaria un'ulteriore ricerca teorica.

Esiste anche una relazione tra lo sviluppo di idee nella fisica di base e nuovi concetti matematici. Per esempio, è stato possibile portare la supergravità al suo livello attuale di raffinatezza perché i matematici, da parte loro, avevano sviluppato algebre non commutative direttamente applicabili alle teorie fisiche. È possibile che una comprensione più profonda del ruolo dello spazio e del tempo nella teoria dei quanti richieda lo sviluppo e l'introduzione di ulteriori concetti matematici; l'interesse attuale per le teorie della gravità con un maggior numero di dimensioni può essere solo un primo passo in questa direzione.

(RI)CREAZIONI AL CALCOLATORE

di A. K. Dewdney

Un bestiario di virus, bachi e altre insidie per la memoria dei calcolatori nella Guerra dei nuclei

Quando nel luglio 1984 uscì il mio articolo sulla Guerra dei nuclei, non mi rendevo conto di quanto serio fosse l'argomento affrontato. La mia descrizione di programmi in linguaggio macchina che si aggirano nella memoria di un calcolatore cercando di distruggersi l'un l'altro ha avuto notevole risonanza. Secondo i resoconti di molti lettori, abbondano gli esempi di bachi, virus e altre creature software annidate in tutti i possibili ambienti di elaborazione. Certe possibilità sono così orribili che esito a riportarle.

Il romanzo francese di spionaggio *Softwar: La Guerre Douce* presenta un'immaginaria situazione geopolitica di questo genere. Gli autori, Thierry Breton e Denis Beneich, imbastiscono un agghiacciante racconto a proposito dell'acquisto, da parte dell'Unione Sovietica, di un supercalcolatore americano. Invece di bloccare la vendita, le autorità americane acconsentono alla transazione mostrando una studiata riluttanza: il calcolatore, infatti, è stato segretamente programmato con una «bomba software». Ufficialmente acquistata per le previsioni meteorologiche sul vasto territorio dell'Unione Sovietica, la macchina, o meglio il suo software, contiene un «grilletto» nascosto: appena il Servizio meteorologico nazionale degli Stati Uniti comunica il rilevamento di una certa temperatura a St. Thomas, nelle Virgin Islands, il programma procede a sovvertire e distruggere tutti i pezzi di software che riesce a trovare nella rete sovietica. Se è vero che sceneggiature di questo genere rappresentano possibilità reali, sono tentato di dire: «Se guerra [*war*] deve essere, che sia almeno dolce [*soft*]». D'altra parte, un dubbio mi viene dalla possibilità di un incidente dovuto al collegamento stretto fra software militare e sistemi di controllo delle armi.

ISTRUZIONE	MNEMONICA	CODICE	ARGOMENTI		SPIEGAZIONE
Sposta	MOV	1	A	B	Sposta il contenuto dell'indirizzo *A* all'indirizzo *B*
Somma	ADD	2	A	B	Somma i contenuti dell'indirizzo *A* e dell'indirizzo *B*
Sottrae	SUB	3	A	B	Sottrae il contenuto dell'indirizzo *A* dall'indirizzo *B*
Salta	JMP	4	A		Trasferisce il controllo all'indirizzo *A*
Salta se zero	JMZ	5	A	B	Trasferisce il controllo all'indirizzo *A* se il contenuto dell'indirizzo *B* è zero
Salta se maggiore	JMG	6	A	B	Trasferisce il controllo all'indirizzo *A* se il contenuto dell'indirizzo *B* è maggiore di zero
Decremento: salta se zero	DJZ	7	A	B	Sottrae 1 dal contenuto dell'indirizzo *B* e trasferisce il controllo all'indirizzo *A* se il contenuto dell'indirizzo *B* diventa zero
Confronta	CMP	8	A	B	Confronta i contenuti degli indirizzi *A* e *B*; se sono diversi, salta l'istruzione successiva
Divide	SPL	9	A		Divide l'esecuzione nell'istruzione successiva e nell'istruzione in *A*
Enunciato di dati	DAT	0		B	Enunciato non eseguibile; *B* è il valore dei dati

Elenco di istruzioni per la Guerra dei nuclei

Prima di passare a descrivere le esperienze avute da vari lettori con programmi ostili, vale la pena di riassumere le caratteristiche principali della Guerra dei nuclei per coloro che non avessero letto l'articolo del luglio 1984.

Due giocatori scrivono ciascuno un programma in un linguaggio di basso livello chiamato REDCODE. I programmi vengono posti in una vasta arena circolare che chiamiamo Nucleo: in realtà nient'altro che una matrice di migliaia di locazioni, in cui l'ultimo indirizzo è contiguo al primo. Ogni istruzione del programma da battaglia occupa una locazione nel Nucleo. Il programma esecutivo MARS (acronimo per *Memory Array Redcode Simulator*, ovvero «simulatore di Redcode nella matrice di memoria») fa girare i programmi da battaglia eseguendo alternativamente un'istruzione dell'uno e una dell'altro, come un semplice sistema a partizione di tempo: i due programmi si attaccano e, a turno, cercano di evitare danni o di riparare quelli subiti. Una semplice modalità d'attacco può essere eseguita per mezzo di istruzioni MOV. Per esempio

MOV # 0 1000

fa sì che il numero 0 sia posto nella locazione il cui indirizzo si trova 1000 locazioni al di là di questa istruzione, cancellando il precedente contenuto di quella locazione. Nel caso lo 0 venisse posto su un'istruzione dell'avversario, anch'essa sarebbe tolta di mezzo e il programma non sarebbe più eseguibile: l'avversario avrebbe perso il gioco.

Dato che nessun calcolatore, personale o *mainframe*, è dotato all'origine di REDCODE e di un'adeguata matrice da battaglia, queste caratteristiche devono essere simulate. È ancora disponibile la traccia per la stesura di un programma di simulazione (può essere richiesta ancora a «Le Scienze», rubrica «(Ri)creazioni al calcolatore», via del Lauro 14, 20131 Milano, inviando 2000 lire, anche in francobolli, per le spese di fotocopia e di spedizione). L'anno scorso, parecchie centinaia di lettori hanno acquistato questa traccia e molti fra loro hanno scritto programmi di gioco per la Guerra dei nuclei.

Ispirandosi a un articolo di L. S. Penrose sui meccanismi che si autoriproducono, apparso su «Scientific American» nel giugno 1959, Frederick G. Stahl di Chesterfield, Missouri, ha creato un universo lineare in miniatura in cui umili creature vivono, si muovono e (in un certo senso) compiono il proprio destino. Scrive Stahl:

«Come nella Guerra dei nuclei, ho isolato un segmento lineare chiuso di memoria principale in cui una creatura era simulata dal linguaggio macchina modificato. La macchina era un IBM Tipo 650 con memoria a tamburo. La creatura era programmata per strisciare lungo il suo universo mangiando cibo (parole diverse da zero) e creando un duplicato di se stessa quando era stato accumulato abbastanza cibo. Come nella Guerra dei nuclei, avevo un programma esecutivo che teneva conto di chi era vivo e distribuiva il tempo d'esecuzione tra le creature viventi. Lo chiamavo «la mano sinistra di Dio». Stahl prosegue analizzando la capacità del suo programma di riprodursi e descrivendo un interessante meccanismo di mutazione: un programma potrebbe subire, durante la copiatura, un piccolo numero di cambiamenti casuali nel suo codice. Tuttavia, riferisce Stahl, «abbandonai questa linea di lavoro dopo una sessione di produzioni in cui un mutante sterile uccise e mangiò l'unica creatura feconda dell'universo. Era chiaro che per ottenere qualche risultato interessante sarebbero state necessarie memorie incredibilmente grandi e tempi di elaborazione lunghissimi.»

```
1   IF PEEK (104) = 134 GOTO 10
2   POKE 104, 134: POKE 134 * 256,0
3   PRINT CHR$(4) "RUN APPLE WORM"
10  HOME : POKE - 16302,0: POKE - 16304,0: POKE 1023,160
20  FOR I = 0 TO 94: READ D: POKE 1024 + I, D: NEXT I
30  POKE - 16368,0
40  IF PEEK ( - 16384)<128 GOTO 40
50  CALL 1024
100 DATA 160,225,200,185,255,3,153,127,4,192,95,208,245,
    160,18,190,76,4,24,189,128,4,105,128,157,128,4,189,129,
    4,105,0,157,129,4,192,13,208,18,238,23,4,173,23,4
200 DATA 141,151,4,206,31,4,173,31,4,141,159,4,136,208,211,
    173,167,4,72,173,176,4,141,167,4,104,141,176,4,76,128,
    4,7,20,25,28,33,46,55,61,65,68,72,75,4,16,40,43,49,52
```

Un baco che vive negli Apple

Una storia analoga riguarda un gioco chiamato Animale, in cui un programma cerca di stabilire a che animale sta pensando un uomo. David D. Clark, del Laboratory for Computer Science del Massachusetts Institute of Technology, scrive che gli impiegati di una certa azienda giocavano con vero ardore ad Animale. Pur non somigliando a un programma di battaglia e nemmeno alle semplici creature di Stahl, Animale aveva la capacità di riprodursi nei corridoi del nucleo sfruttando gli sforzi del programmatore di potenziare una caratteristica chiave del gioco: quando sbaglia nell'indovinare l'animale che il giocatore umano ha in mente, il programma chiede a quest'ultimo di suggerire una domanda che esso potrebbe porre per migliorare le sue prestazioni future. Questa caratteristica, prosegue Clark, portò il programmatore a inventare un trucco per assicurarsi che ognuno avesse sempre la stessa versione di Animale.

«Su un sistema di elaborazione antiquato, privo di una struttura di catalogo condivisa e privo anche di mezzi di protezione, un programmatore inventò un modo molto originale per rendere disponibile il gioco a più utenti. Una versione del gioco esisteva nel catalogo degli archivi di un utente; quando questi giocava, il programma produceva una copia di se stesso in un altro catalogo di archivi. Se quel catalogo conteneva già una copia del gioco, la vecchia versione veniva cancellata, così che il comportamento del gioco cambiava in modo inatteso per il giocatore. Se quel catalogo non conteneva in precedenza una versione di Animale, un altro utente si trovava ad avere a disposizione il gioco.»

Clark ricorda che Animale era un gioco talmente popolare che, alla fine, tutti i cataloghi del sistema dell'azienda ne contenevano una copia. «Quando poi degli impiegati dell'azienda venivano trasferiti ad altre divisioni... portavano con sé anche Animale, che così si diffuse di macchina in macchina all'interno dell'azienda.» La situazione non sarebbe mai diventata seria se non fosse stato che tutte quelle copie del gioco, per altri versi innocue, cominciarono a intasare la memoria su disco. Solo quando qualcuno inventò una versione più «virulenta» del gioco, la situazione poté tornare sotto controllo. Quando la nuova versione di Animale veniva giocata, essa si copiava in altri cataloghi non una ma due volte. Dandogli abbastanza tempo, si pensava, questo programma avrebbe alla fine cancellato tutte le vecchie versioni di Animale. Dopo un anno, al raggiungimento di una data prefissata, in ogni copia del nuovo programma Animale si innescò un nuovo meccanismo. «Invece di replicarsi due volte, ora esso giocava una partita finale, augurava "arrivederci" all'utente e poi si cancellava. Così Animale venne espulso dal sistema.»

Una volta Ruth Lewart di Holmdel, New Jersey, creò un mostro (per così dire) senza neanche scrivere un programma. Mentre lavorava, sul sistema a partizione di tempo della sua azienda, alla preparazione di una versione dimostrativa di un programma didattico, decise di produrre una copia di riserva su un altro sistema a partizione di tempo. Quando il sistema originale cominciò ad apparire lento, riferisce, «inserii il sistema ausiliario, che era molto sensibile, per tre minuti interi, durante i quali non ci fu alcuna risposta e il caos più completo regnava sullo schermo del mio terminale grafico. Nessuno era più in grado di inserirsi o di uscire dal sistema. La conclusione che si poteva trarre era una sola: in qualche modo la colpa era del mio programma! Nonostante il panico, mi resi improvvisamente conto di aver usato una "e" commerciale (&) come carattere separatore di campo del terminale. Ma la & era anche il carattere usato dal sistema per generare un processo di fondo. Alla prima lettura dallo schermo, il calcolatore deve aver intercettato le & indirizzate al terminale e deve aver generato un gran numero di processi, ciascuno dei quali, a sua volta, generò altri processi, e così via all'infinito.» Una telefonata frenetica informò un responsabile di sistema dell'origine del disturbo e il calcolatore *mainframe* venne spento e fatto ripartire. Inutile dire che Lewart cambiò la & in un altro carattere e il suo programma «da allora ha sempre funzionato felicemente».

Anche se i programmi per la Guerra dei nuclei non vengono generati in questo modo, copie aggiuntive possono aumentare le loro capacità di sopravvivenza. Parecchi lettori hanno proposto la realizzazione di tre copie del programma, in modo che la copia in esecuzione possa utilizzare le altre due per stabilire se qualche sua istruzione è sbagliata. Il programma in esecuzione potrebbe sostituire un'istruzione difettosa con una idonea a garantire la sopravvivenza. Un'idea analoga sta alla base di Scavenger (Spazzino), un programma ideato per proteggere da errori gli archivi di una memoria di massa quando si preparano copie di riserva su nastro magnetico. Arthur Hudson, che vive a Newton nel Massachusetts (e lavora per un'altra azienda che non citiamo), scrive: «Chiunque abbia usato spesso il nastro magnetico si è trovato assediato da una forza aliena chiamata Legge delle probabilità composte.» Hudson prosegue citando vari errori connessi con l'uso di nastri e dimostra che, sebbene ogni tipo di errore abbia una probabilità relativamente piccola di prodursi, la probabilità che se ne verifichi almeno uno è

DAT		0	/puntatore all'indirizzo da proteggere
ADD	#1	—1	/incrementa l'indirizzo della protezione
PCT	@—2		/protegge l'indirizzo
CMP	#102	—3	/se tutte e 102 le istruzioni sono protette...
JMP	2		/... allora lascia il ciclo
JMP	—4		altrimenti riprendilo

⋮

CORPO DEL PROGRAMMA DI BATTAGLIA PRINCIPALE

⋮

Questo ciclo protegge i combattenti da bombe vaganti

spiacevolmente alta. Poi continua così:

«Niente paura. Scavenger è con voi: se gli affidate un archivio di una memoria di massa, copierà l'archivio su tre nastri magnetici senza seccarvi con banali dettagli di gestione. Anche se il calcolatore cade in errore logico (come avveniva parecchie volte al giorno) di solito non verrà distrutta la registrazione dell'esecuzione; quando il calcolatore viene riattivato, tutti i bachi presenti nella registrazione si metteranno a loro volta in funzione. Alla scrittura di ciascun nastro presiede un compito distinto, sotto il coordinamento di un programma principale.»

Chi possiede un calcolatore Apple dovrebbe guardarsi da uno spregevole programmino chiamato Apple Worm (letteralmente Baco della mela), creato da Jim Hauser e William R. Buckley della California Polytechnic State University a San Luis Obispo. Scritto per l'Apple II in linguaggio di assemblatore del microelaboratore 6502, questa specie di baco si riproduce nel corso di un allegro viaggetto attraverso l'Apple ospite. All'inizio si carica un particolare programma BASIC *(si veda l'illustrazione della pagina precedente)*, il quale a sua volta carica il baco nella parte bassa della memoria (quella con indirizzi bassi). Il programma BASIC occupa invece la parte alta della memoria.

«Dato che il Baco viene caricato in una delle aree grafiche della macchina, si può osservare il Baco che inizia il suo attacco a capofitto (o, per meglio dire, a codafitta) nella memoria alta... Una volta che il Baco ha lasciato la finestra grafica... si può aspettare che abbia cancellato la parte alta della memoria (compreso il programma BASIC) e vada a scontrarsi con le ROM di sistema.»

Hauser e Buckley hanno in programma di pubblicare, in un futuro non lontano, una raccolta di bachi. Hanno progettato un Worm Operating System (Sistema Operativo Baco) e hanno perfino scritto un videogioco in cui Baco ricopre uno dei ruoli principali.

Un'altra minaccia software viene proposta da Roberto Cerruti e Marco Morocutti di Brescia. Prendendo spunto dalla traduzione italiana dell'articolo sulla Guerra dei nuclei apparsa su «Le Scienze», hanno pensato a un modo per infestare l'Apple II, non con un baco ma con un virus. Scrive Cerruti:

«Marco pensò di scrivere un programma capace di passare da un calcolatore all'altro, come un virus, e di "infettare" in questo modo altri Apple. Non fummo però in grado di idearlo finché non mi resi conto che il programma doveva infettare i dischetti e usare il calcolatore solo come mezzo per migrare da un disco all'altro. Fu così che il nostro virus cominciò a prendere forma.

«Come si sa, ogni dischetto Apple contiene una copia del DOS (il sistema operativo per dischi), che viene lanciato dal calcolatore non appena riceve l'alimentazione. Il virus era una modificazione di questo DOS, che a ogni operazione di scrittura verificava la propria presenza sul disco e, in caso negativo, modificava il DOS sul disco, copiandosi così su ogni dischetto messo nell'unità di lettura e registrazione dopo la prima accensione. Pensammo che, se avessimo installato un simile DOS su alcuni dischi usati nel maggiore negozio di calcolatori della nostra città, Brescia, in breve avremmo provocato la diffusione di un'epidemia in tutta la città.

«Ma era una vera epidemia, con virus così innocui? No, i nostri virus dovevano essere maligni! Decidemmo, quindi, che dopo 16 cicli di autoriproduzione, contati sul disco stesso, il programma dovesse decidere di reinizializzare il disco subito dopo il lancio. A quel punto era chiara la terribile perversità della nostra idea e decidemmo di non metterla in pratica né di parlarne con alcuno.»

Cerruti e Morocutti sono stati gentili. In un calcolatore personale, il sistema operativo per dischi è l'arbitro ultimo del destino dei programmi, dei dati e di tutto il resto. Nello schema appena descritto, il sistema operativo infetto cancella il disco da cui è stato lanciato e quindi non può più essere caricato se non da un altro disco. Il DOS contagiato potrebbe anche far apparire periodicamente un messaggio irritante del tipo:

IL VOSTRO DISCO
VI SFUGGE?
È ora di rivolgersi al
DOTTOR DOS
disponibile su disco nel
più vicino negozio di calcolatori

L'infezione virale descritta si è già verificata su piccola scala. Richard Skrenta, Jr., studente di una scuola superiore di Pittsburgh, scrisse un programma di questo genere. Invece di cancellare dischi o visualizzare avvisi, questa forma di infezione provocava la comparsa di subdoli errori nel sistema operativo.

«Tutto questo sembra molto infantile», scrive Skrenta, ma «ahimè! Non son più stato capace di liberarmi del mio flagello elettronico. Ha contagiato tutti i miei dischi e i dischi dei miei amici. È riuscito addirittura ad arrivare ai dischi con i programmi per i grafici di funzione del mio professore di matematica.» Skrenta ha inventato un programma per distruggere il virus, ma non si è mai dimostrato efficace quanto il virus stesso.

Quanto detto implica un problema interessante, e sarei privo di fantasia e irresponsabile se non lo ponessi: descrivere, in una pagina o meno, DOTTOR DOS, un programma su disco che in qualche modo sappia soffocare epidemie elettroniche di questo genere. Molti dischi usati da un calcolatore personale contengono copie del suo DOS. Quando viene fatto partire, il calcolatore ottiene dal disco la sua copia del DOS, che rimarrà operante anche quando vengono fatti girare altri dischi, sia pure contenenti anch'essi copie del DOS. Se è infetto, il DOS attivo può alterare le altre copie del DOS o addirittura sostituirle con copie di se stesso. Come ci si può opporre a questa virulenza?

Nella versione iniziale della Guerra dei nuclei, la difficoltà principale consisteva nel fornire al programma di battaglia A i mezzi per proteggersi dai colpi vaganti del programma di battaglia B. Se questo tipo di protezione riusciva a essere più o meno efficace, l'evoluzione del gioco portava al livello successivo, in cui i programmi sarebbero stati costretti a cercarsi l'un l'altro e a sviluppare attacchi concentrati.

Nel tentativo di garantire questa protezione, nell'articolo del luglio 1984 suggerivo l'istruzione

PCT *A*

dove *A* è l'indirizzo relativo (diretto o indiretto) di un'istruzione che deve essere protetta. Un tentativo di cambiare il contenuto di quell'indirizzo sarebbe bloccato da MARS, il sistema supervisore del gioco. Il tentativo successivo, però, avrebbe avuto successo. Mi sembrava che, impiegando un semplice ciclo, un programma di battaglia potesse proteggere tutte le proprie istruzioni da bombe vaganti abbastanza a lungo da riuscire a lanciare una sonda indisturbata verso l'altro programma. La figura di questa pagina mostra in forma schematica un programma ad autoprotezione di questo genere. Il ciclo di protezione è formato da sei istruzioni, quattro delle quali eseguite a ogni passaggio nel ciclo. Così, un

programma di battaglia di n istruzioni (ciclo incluso) richiederebbe circa $4 \times n$ esecuzioni per avere una protezione completa da un singolo colpo. Questo scudo non è una gran protezione contro un programma dwarf che lanci due colpi contro ogni locazione.

Esiste un altro uso di questa istruzione, non previsto nel precedente articolo sulla Guerra dei nuclei. Stephen Peters di Timaru, Nuova Zelanda, e Mark A. Durham di Winston-Salem, North Carolina, l'uno indipendentemente dall'altro, hanno pensato di usare PCT in modo offensivo. Un programma chiamato TRAP-DWARF innalza uno sbarramento di zeri nel solito modo ma poi protegge ogni deposito contro la sovrascrittura. Questo significa che un programma nemico incauto può cadere in una di queste trappole mentre si trascrive in una nuova area. L'istruzione indirizzata alla locazione occupata da uno zero protetto non avrebbe ovviamente effetto su quella locazione. Quando, in seguito, l'esecuzione raggiunge quell'indirizzo, il nuovo programma muore perché 0 non è un'istruzione eseguibile. Varrà forse la pena di includere PCT in qualche futura versione della Guerra dei nuclei, ma per ora la terrò da parte per amore di semplicità, sigillo del progettista del gioco.

Tra le altre idee suggerite dai lettori ci sono la matrice di nuclei bidimensionale proposta da Robert Norton di Madison, Wisconsin, e la regola di limitazione di campo avanzata da William J. Mitchell del dipartimento di matematica della Pennsylvania State University. L'idea di Norton si spiega da sé; la proposta di Mitchell, invece, richiede un approfondimento. Essa consiste nel consentire a ogni programma di battaglia di modificare il contenuto di qualsiasi locazione che non disti più di un certo numero prestabilito di indirizzi. Una regola di questo genere impedisce automaticamente a DWARF di fare danni al di fuori di questo intorno. La regola ha anche molti altri effetti, tra cui quello di sottolineare notevolmente il movimento: in quale altro modo un programma di battaglia potrebbe raggiungere il campo di un avversario? La regola ha molti pregi e spero che qualcuno dei molti lettori che possiedono un sistema per la Guerra dei nuclei le voglia dedicare l'ulteriore approfondimento che merita.

Norton propone anche che, in una battaglia della Guerra dei nuclei, a ogni contendente sia consentita più di una esecuzione. La stessa idea è venuta a molti altri lettori e ho deciso di accettare il suggerimento, così che ora la Guerra dei nuclei assume un carattere di grande apertura che prima mancava.

Il gioco si modifica aggiungendo la seguente istruzione, chiamata scissione, al listato ufficiale della Guerra dei nuclei *(si veda l'illustrazione di pagina 106)*.

```
SPL A
```

Quando l'esecuzione raggiunge questo punto, si divide in due parti. l'istruzione che segue SPL e quella distante A indirizzi. Questo consente immediatamente a ogni giocatore della Guerra dei nuclei di far girare più programmi alla volta, quindi è necessario definire il modo in cui MARS assegnerà queste esecuzioni. Sotto questo profilo, esistono due diverse possibilità.

Per illustrarle, supponiamo che un giocatore abbia i programmi A_1, A_2 e A_3, mentre l'altro giocatore ha i programmi B_1 e B_2. Un'alternativa è far girare tutti i programmi del primo giocatore, seguiti da quelli del secondo giocatore. L'ordine dell'esecuzione sarebbe così A_1, A_2, A_3 e poi B_1 e B_2, e il ciclo si ripeterebbe indefinitamente. La seconda possibilità è alternare i programmi dei due giocatori. In questo caso la successione sarebbe A_1, B_1, A_2, B_2, A_3, B_1 e così via. I due schemi sono molto diversi. Il primo mette l'accento su una proliferazione illimitata e sembra quindi limitare il ruolo dell'intelligenza nel gioco. Nel secondo, invece, quanto maggiore è il numero di programmi fatti girare dai due giocatori, tanto minore è il numero di volte che ciascuno di essi sarà eseguito. In questo contesto sembra appropriata una legge dei ritorni decrescenti, quindi ho adottato il secondo schema. Scopo del gioco, in ogni caso, è provocare l'arresto di tutti i programmi nemici.

La nuova istruzione è ricca di possibilità creative. Per illustrarne una delle più modeste, ecco un programma di battaglia chiamato IMP GUN:

```
SPL 2
JMP -1
MOV 0 1
```

Consideriamo quello che avviene quando l'esecuzione arriva per la prima volta alla sommità di questo programma. SPL 2 significa che in seguito saranno assegnate a questo programma due esecuzioni: saranno eseguite sia JMP −1 sia MOV 0 1. La prima istruzione farà sì che il programma rientri nel ciclo e la seconda mette in movimento un IMP. L'IMP si muoverà verso il basso, naturalmente, dato che l'obiettivo del comando MOV sarà sempre l'indirizzo successivo, come indicato dall'1 (positivo). L'IMP così viene generato a ogni ciclo del programma e un flusso senza fine di esecuzioni di IMP scorre attraverso il nucleo deciso a distruggere i programmi nemici. A prima vista, può sembrare che non ci sia alcuna difesa possibile contro un simile esercito di IMP; in realtà una c'è. Bisogna mettere in gioco IMP PIT, un programma ancora più semplice, attivato da un comando SPT inserito in un insieme più esteso di istruzioni volte a proteggere il suo fianco superiore:

```
MOV # 0 -1
JMP -1
```

A ogni esecuzione, IMP PIT pone uno zero subito sopra di sé, nella speranza di distruggere un IMP in arrivo. Qui è fondamentale la regola di esecuzione-assegnazione. Se IMP GUN appartiene ad A e IMP PIT appartiene a B, allora A richiede n mosse per eseguire n IMP; solo un IMP può arrivare alla locazione subito sopra l'IMP PIT. A parità di altre condizioni, B deve eseguire IMP PIT solo una volta per eliminare un IMP in arrivo.

Nella versione allargata del gioco della Guerra dei nuclei, si immagina che ogni contendente generi e metta in campo piccoli eserciti di programmi formulati singolarmente per individuare, attaccare, proteggere e anche riparare. Numerose sottigliezze, come quella proposta da John McLean di Washington, D.C., richiedono un'analisi ulteriore. McLean immagina un programma trappola specializzato, che sistema comandi JMP in vari indirizzi in tutta la matrice del nucleo nella speranza di far approdare un comando JMP all'interno di un programma nemico. Ogni JMP collocato in questo modo trasferirebbe l'esecuzione del programma nemico al programma trappola, provocandone per così dire il passaggio al nemico.

Dalla mischia provocata dai programmi di battaglia emerge un problema importante, che ha bisogno di soluzione. Che cosa impedisce a un programma di battaglia di uno dei contendenti di attaccare i suoi colleghi? Appare necessario un sistema di ricognizione.

Tra i molti lettori che hanno costruito sistemi per la Guerra dei nuclei meritano una citazione particolare: Chan Godfrey di Wilton, Connecticut, Graeme R. McRae di Monmouth Junction, New Jersey, e Mike Rosing di Littleton, Colorado, perché hanno messo particolare cura nel definire e documentare i loro progetti. Mi piacerebbe in particolare rendere disponibili ai lettori i documenti di Rosing, ma ho un'altra idea migliore che include questa possibilità e risolve anche altri problemi di comunicazione. Se qualche lettore con un sistema per la Guerra dei nuclei già funzionante si offrirà come direttore di una rete di Guerra dei nuclei, allora si potranno comunicare a tutti gli utenti della Guerra dei nuclei una documentazione dei vari sistemi, proposte di regole, programmi interessanti e battaglie. Un volontario sarà scelto come direttore; gli altri volontari potrebbero dar vita, secondo i loro interessi, a un bollettino, a un comitato per le regole, e così via. In un articolo futuro darò il nome e l'indirizzo del direttore di rete.

Continuano ad arrivare numerosi resoconti di lettori che hanno giocato con l'ecologia del pianeta Wa-Tor (si vedano le «(Ri)creazioni al calcolatore» del febbraio 1985); potrò quindi esaminare solo poche fra le molte esperienze descritte. In linea generale, la scelta dei giusti parametri ha prodotto grosse fluttuazioni nelle popolazioni degli squali e dei pesci. Alcuni lettori, desiderosi di rendere Wa-Tor più simile alla Terra, hanno aggiunto particolari caratteristiche ai loro programmi. Il gioco, in effetti, invita a una sua complicazione, che è certamente benvenuta. L'introduzione di un sistema variante, però, ha il grosso svantaggio (a parità di altri fattori) di rendere ardui i confronti con il sistema standard.

Costruttori di un sistema iniziale sono stati Jean H. Anderson di Lauderdale, Minnesota, Stephen R. Berggren di Satellite Beach, Florida, Milton Boyd di Amherst, New Hampshire, J. Connett di Minneapolis, Minnesota, Edgar F. Couda di Park Ridge, Illinois, Jim Lemon di El Segundo, California, e Kenneth D. Wright di Grayling, Michigan.

Tra le questioni che questi e altri lettori si sono trovati ad affrontare c'era la durata della sopravvivenza. Chiaramente, non c'è alcun problema per le popolazioni eterne, ma sarebbe utile avere una misura delle sceneggiature che non sono eterne. Come rileva Stevens, misurare con i crononi può essere fuorviante quando si scelgono la durata delle estensioni di vita e i tempi di riproduzione per gli squali. Anche la misurazione per cicli solleva problemi: che cos'è un ciclo? Stevens fa la divertente osservazione che se gli squali e i pesci sopravvivono a un sufficiente numero di ripetizioni del ciclo base con numeri casuali, si ripeterà una configurazione precedente, in accordo con il ciclo, e da lì in avanti alle popolazioni sarà ovviamente garantita la vita eterna.

Un gran numero di lettori, tra cui David Emanuel di Oak Brook, Illinois, Richard G. Fizell di Fort Washington, Maryland, e John S. Lew dello IBM Thomas J. Watson Research Center di Yorktown, New York, hanno descritto teorie moderne utili all'analisi di Wa-Tor. Non è stata ancora detta l'ultima parola a proposito del dilemma se le matrici stocastiche ci metteranno in grado di derivare specifiche probabilità di sopravvivenza da combinazioni arbitrarie di parametri o no. È interessante notare, però, che le equazioni di Lotka-Volterra (dalla loro formulazione nel 1931) sono state elaborate in modo da prendere in considerazione la diffusione come fattore che riguarda sia il predatore sia la preda. La diffusione fa assumere forme più complesse alle soluzioni di Lotka-Volterra, che di solito variano con regolarità. Una nota storica di Lew ci precisa come Alfred J. Lotka fosse un matematico americano il quale, un decennio prima, aveva formulato la stessa equazione di Volterra.

Boyd ha sfruttato un diagramma di fase per analizzare la dinamica delle popolazioni di squali e altri pesci. A ogni istante t, si riportano in grafico il numero x dei pesci e quello y degli squali come coordinate di un singolo punto. Man mano che il tempo avanza e le popolazioni seguono il loro ciclo, il punto descrive un'orbita erratica intorno a un occhio, o centro, fisso. Boyd ha usato questa tecnica per studiare l'effetto delle dimensioni dell'oceano sulla sopravvivenza e scrive che «per i mondi più rettangolari, le orbite persero il loro occhio, le traiettorie divennero più nervose, per diventare infine percorsi casuali». Evidentemente, sono preferibili oceani quadrati.

Tra le innovazioni introdotte dai lettori possiamo annoverare una forza vitale degli squali, mutazioni, doppie popolazioni di pesci e plancton. Nel mio articolo di febbraio avevo trascurato di dire che i pesci comuni di Wa-Tor si cibano di un plancton oceanico sparso ovunque in abbondanza. Lemon ha reso esplicita questa caratteristica facendo in modo di mettere plancton in ogni punto non occupato da uno squalo o da un altro pesce. Il plancton prolifica in punti altrimenti vuoti e ha con i pesci comuni la stessa relazione che i pesci comuni hanno con gli squali. Anche in questo caso esistono popolazioni eterne.

Gli squali di Couda guadagnano o perdono punti di forza vitale a seconda di quanto mangiano. Possono così sopravvivere senza cibo molto più a lungo dei semplici squali dello Wa-Tor standard. Couda (come molti altri programmatori di Wa-Tor) ha inviato diagrammi che sono notevolmente simili a quelli che si ottengono in base ai dati della Hudson's Bay Company.

Connett utilizza due specie di pesci. Una è la varietà standard di Wa-Tor; l'altra si riproduce sempre in qualsiasi punto vuoto a sud o a est. A causa della sua tendenza alla mobilità, la seconda specie spesso sopravvive alla prima. Rudy Iwasako di Sacramento, California, ha proposto di attribuire agli squali e agli altri pesci caratteristiche di dimensione, velocità e agilità, sottoposte a controllo genetico. Berggren ha scritto il suo sistema, chiamato EVOLVE, due anni fa. Esso è simile a WATOR, ma lascia che gli animali si evolvano secondo le pressioni ambientali. A giudizio di Berggren, le popolazioni arriverebbero a un equilibrio favorevole alla sopravvivenza a lungo termine.

Nessuno è riuscito a risolvere il problema dell'inseguimento toroidale. Rivelerò solo metà della soluzione, in modo da riservare ai lettori il piacere di trovare l'altra metà. Si ricordi che, a ogni turno, il pesce si muove e poi si muovono i due squali. Come in Wa-Tor, non è consentito rimanere nello stesso punto. Ogni raggio segue una diagonale e gira intorno al toro, ricongiungendosi presto o tardi con se stesso. Quando entrambi gli squali occupano una coppia di raggi opposti, non importa in che modo il pesce si muove: uno squalo insegue a distanza costante e l'altro squalo si avvicina. Il pesce è condannato. Lascio ai lettori scoprire in che modo gli squali, per così dire, vadano alla caccia dei raggi.